Report No. FAA-RD-70-70

A STUDY OF AIR TRAFFIC CONTROL SYSTEM CAPACITY

Gordon Raisbeck, Bernard O. Koopman, Simon F. Lister, and Asha S. Kapadia

ARTHUR D. LITTLE, INC.
Cambridge, Massachusetts

October 1970

Interim Report

September 1969 to August 1970

Availability is unlimited. Document may be released to the Clearinghouse for Federal Scientific and Technical Information, Springfield, Virginia 22151, for sale to the public.

Prepared for

FEDERAL AVIATION ADMINISTRATION
Systems Research and Development Service
Washington, D.C. 20590

162

1. Report No. FAA-RD-70-70

2. Government Accession No.

3. Recipient's Catalog No

4. Title and Subtitle

A STUDY OF
AIR TRAFFIC CONTROL SYSTEM CAPACITY

5. Report Date
October 1970

6. Performing Organization Code
71860

7. Author(s)
G. Raisbeck, B. O. Koopman, S. F. Lister, A. S. Kapadia

8. Performing Organization Report No.

9. Performing Organization Name and Address

Arthur D. Little, Inc.
Cambridge, Mass. 02140

10. Work Unit No

11. Contract or Grant No
FA70WA-2141

12. Sponsoring Agency Name and Address

Systems Research and Development Service
Federal Aviation Agency
Dept. of Transportation, Washington, D. C.

13. Type of Report and Period Covered
Interim Report
(Sept. 1969 – Aug. 1970)

14. Sponsoring Agency Code

15. Supplementary Notes

None

16. Abstract

This report describes the work on capacity measurement methodology for the air traffic control (ATC) system that Arthur D. Little, Inc., has completed in the initial year of a projected 5-year program. The long-range objective of the program is to develop tools and techniques to define, measure, and predict the quantitative capacity of an air traffic control system, which can then be used in analytical studies in support of long-range plans, management decisions, and system perfo mance evaluations.

A first approximation to a functional description of the present air traffic control system, largely free of descriptions of equipment or methods now used to implement the functions, has been prepared. Several approaches to the quantitative use of measures of safety as system variables have been formulated. The mathematical theory of time-dependent queues was found to be applicable to air traffic control system capacity problems, although previous workers in this field have made little use of it. Some elementary applications, chiefly illustrative of methods, have been made.

In the context of functional descriptions of air traffic control, some preliminary attempts have been made to make the concepts of air traffic control system capacity precise. We have concluded:

1. A ctional description of air traffic control which is applicable to a wide range of system concepts, inc ng the present system and many suggested variations and alternatives, is feasible;
2. The athematical theory of time-varying queues is an important, useful tool;
3. The quantitative study of safety is highly relevant to ATC capacity, but must be studied indirectly rather than through accident statistics; and
4. Simple, straightforward definitions of concepts, such as capacity, demand, delay, and safety, are unlikely, but the need for them can be satisfied by families of less comprehensive terms.

17. Key Words

- ● Air Traffic Control
- ● Capacity
- ● Methodology
- ● Non-Stationary Queue
- ● Demand
- ● Delay
- ● Safety
- ● System Modelling

18. Distribution Statement

Availability is unlimited. Document may be released to the Clearinghouse for Federal Scientific and Technical Information, Springfield, Virginia 22151, for sale to the public.

19. Security Classif. (of this report)
None

20. Security Classif. (of this page)
None

21. No. of Pages
165

22. Price
$3.00 (hard copy)
$1.50 (microfiche)

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF FIGURES (Continued)

Figure No. Page

1. SUMMARY

The number of aircraft flying today could not operate safely and economically without air traffic control. As the number of aircraft operations increases, the variety of air transportation services is extended, and the nation's dependence on air transportation grows, more and better air traffic control is needed But to support the plans, decisions, and actions made to expand and improve the air traffic control system requires analytical tools. Yet, at present, there is no accepted definition of, or means to measure, the capacity of an air traffic control system to function at an acceptable level of service in response to the challenge of a stated traffic demand.

The purpose of this air traffic control system study is to establish measures of the system's effectiveness in performing its functions and to examine its operation, properties, and reactions to various conditions and requirements, so that the effects of proposed changes in equipment, methods of operation, or imposed demands can be foreseen and expressed in terms of these measures of effectiveness. This report describes the first year of activity completed by Arthur D. Little, Inc., in a program expected to last several years.

In this report, we illustrate concepts of capacity, delay, and demand at an air terminal with a simple fluid-flow analogy. Capacity is always found to be a bound of an amount or rate, under given conditions constraining operation, beyond which the quality of service is degraded to an unacceptable level. A complete definition of capacity always requires a statement of operating constraints and of the nature and threshold of service degradation. Multiple meanings have also been attributed to the terms demand and delay, but these concepts have also been restricted and clarified through the use of a fluid–flow analogy.

In a study of the capacity of an air traffic control system, a quantitative concept of safety is especially important: first, because loss of human life is the major service degradation penalty in ou air transportation system, and, second, because other service degradations are traded in order to exceed the threshold of service quality which implicitly defines capacity. In this report, we suggest a number of definitions and measures of safety, including one which has not been used before in air transportation system analysis; viz., the probability of fatality per hour of exposure of the subject. The relative merits of these definitions are illustrated in a number of situations. The new definition is shown to be particularly apt in an analysis of socially acceptable levels of safety.

Analysis of some recent accident records shows that the incidence of fatal aircraft accidents must always remain below the level where statistical analysis can provide useful and timely criteria for air traffic control planning and management decisions. What is required is an indirect measurement of safety, based on some theoretical model of how accidents occur. We have illustrated this factor by showing how to estimate the probability of mid-air collision from a measurement of the distance of the closest approach in near encounters.

Of a number of different ways used to describe the air traffic control system as a whole, we found that the one most suitable for capacity analysis was a functional description; i.e., one which describes the elements of the air traffic control system – and the air transportation system of which it is a part – in terms of the objectives and functions used to achieve these objectives, rather than in terms of the equipment and other means used to carry them out. A description in terms of functions facilitates quantitative comparisons of ATC alternatives, using different equipment, methods, and procedures. Description in terms of goals and purposes makes it easy to show how the benefits of air traffic control accrue.

The air traffic control system is an information subsystem embedded in a transportation system devoted to the physical movement of vehicles and their passengers and cargo from place to place. As an information subsystem, the air traffic control system itself has capacity limitations and operating degradations, but the quality of service which determines whether the capacity is being approached is the quality of air transportation service, not the quality of air traffic control service. It is therefore necessary to complete a functional analysis of the whole air transportation system, referring judgments of system performance to the primary demand for moving people and cargo.

A functional analysis is easily stratified into a hierarchy of larger units each containing smaller units. The means by which the functions of one unit are achieved become the goals of the subsidiary units. The overall air traffic control objective of ensuring safe and efficient use of the national air space by military as well as civil aviation and fostering civil aviation and air commerce is satisfied, in part, by means such as navigation, separation, and regulation. Each of these can be amplified, and further subdivided.

The mathematical theory of time-dependent queues has a number of applications to air traffic control system capacity, but available mathematical resources have not been turned to this purpose. We have identified a class of time-dependent queueing problems with periodic demand and service functions, such as one might use to represent diurnal variations in demand, and demonstrated some general properties of their solution. In particular, we find that a large class of such problems admits of a unique and stable periodic solution.

With the aid of machine computation, we have calculated some of the statistics of a number of time-varying single queues to illustrate dynamic properties not accessible by steady-state analysis. We have also formulated differential equations for double queues with several priority rules.

Those parts of the air transportation system whose form is determined by considerations other than air traffic control, or which may simply be given as inputs or demands, such as the aircraft and the existing air terminals, need not be abstracted in terms of their functions alone. Within the air traffic control system, analyses based on information flow and on interrelationships among control loops help in finding characterization of parts of the system wnich are not peculiar to a particular equipment embodiment.

2. INTRODUCTION

2.1 BACKGROUND

The demand for air transportation is growing rapidly. Year by year we see increases in the number of passengers and the amount of cargo moved by air, the distances over which they are moved, and the speed with which they are transported. Not only is the number of commercial aircraft in use increasing, but the new aircraft are larger than their predecessors. Because of higher aircraft speeds, the average number of flights per hour of operation is increasing, even though the average distance from takeoff to landing is increasing. The number of general-aviation aircraft is increasing rapidly also, and new kinds of air transportation services are being projected for the time when supersonic transports and STOL craft are developed.

Delays and cancellations are already being felt at high-density terminals. Unless improvements are made in the operation of our air transportation system, such degradations of service will increase. Terminals that are already congested will become even more so; and at terminals where such degradations have been negligible, they will become significant. Finally, the service degradations will even spread to large areas in densely populated parts of the country that are far beyond what we now consider to be terminal areas.

We could not fly the number of aircraft operations or move the amount of cargo or the number of passengers which we do today at socially and economically acceptable levels of safety without constraining and ordering the flow of traffic. The number of aircraft which can safely fly from their respective starting points to their intended destinations is greatly increased by standardizing their flight paths and procedures and limiting their freedom of action in ways which add little cost and subtract little value from the service rendered. The control of aircraft operations to achieve safe, efficient transit is the function of air traffic control.

There are a number of reasons why the burden on the air traffic control system is growing even faster than the amount of passenger and cargo flow or the number of aircraft operations:

1. Some air traffic control problems, such as separation control, grow in proportion to the number of (combinatorial) pairs of aircraft, and

Preceding page blank

hence with the square of the total number of aircraft. This is aggravated by the increasing significance of interactions between IFR and VFR aircraft.

2. Although most general aviation aircraft have operated by visual flight rules and made minimal demands for service upon the air traffic control system, the proportion equipped with instruments and flying by instrument flight rules is rapidly increasing.

3. There has been a degree of uniformity in flight dynamics and operating characteristics because of the formerly narrow range of physical and engineering characteristics of aircraft. The landing speed, cruising speed, turning radius, and desirable altitude for operation of aircraft of similar size and propulsion systems fall within a small range. With the development of very large aircraft, supersonic transports, STOL and VTOL aircraft, and other special and extreme types, however, the variety of different flight characteristics which must be accommodated by the air traffic control system is increasing. As the range of operating characteristics of aircraft increases, so does the difficulty of providing an air traffic control system which is compatible with all of them.

4. As air transportation becomes more of a necessity and less of a luxury for our national life style, the cost of doing without air transportation services becomes greater. The option of curtailing operations during unfavorable conditions is discouraged, and incentives are added for sustaining operations at night, in poor weather, and during conditions of low visibility. All of these factors make it harder to provide satisfactory air traffic control.

The Federal Aviation Administration is under pressure to provide more and better air traffic control services. In addition to duplicating equipment and adding personnel to provide more of the same kind of service that is presently available, the FAA is constantly introducing new kinds of equipment, new systems concepts, new air traffic control functions, and new methods of organizing air traffic control. To plan and carry out such innovations, decisions on complex issues must be made in the presence of uncertainty. Accelerating the decision-making processes and improving their accuracy may lead to very large benefits. The FAA is therefore seeking to improve its tools for rational decision-making.

Attempts to make such decisions rationally lead repeatedly to questions about the quantitative relationship between the amount of air transportation supplied and the amount of air traffic control services used to support it. In one form or another, we must answer the following questions:

- Given a set of air traffic control equipment, personnel, and methods, how much traffic can be handled before service is degraded below a certain level?

- In anticipation of a certain demand for air transportation services, what amount of air traffic control equipment, personnel, and services should be provided to fulfill the demand at an acceptable level of service?

- Given a well-ordered set of demands and two or more possible systems or configurations of air traffic control, which configuration will enable the larger demand to be met at a given level of service?

However, by using the word capacity – acknowledging that the term *"capacity of an air traffic control system"* has not yet been precisely defined – we can simplify the cumbersome wording of these questions to read as follows:

- What is the capacity of a particular air traffic control system?

- How much should an air traffic control system be expanded to achieve a particular capacity?

- Which of two (or more) air traffic control systems has the greater capacity?

At present, there is no accepted definition of the capacity of an air traffic control system or one of its subsystems, nor is there an analytical means by which the capacity of the system can be measured. There is no way, other than trial and error, to find whether system improvements intended to increase the capacity are keeping pace with increasing demands, or whether parts of the system have unused capacity, that is, capacity in excess of any demand that has yet been imposed. Until these are developed, judgments about air traffic control needs, benefits, and costs cannot be made on a systematic basis.

2.2 PURPOSE AND SCOPE

The purpose of this air traffic control system study is to establish measures of the system's effectiveness in performing its functions, and to examine its operation, properties, and reactions to various conditions and requirements, so that the effects of proposed changes in equipment, methods of operations, or imposed demands can be foreseen and expressed in terms of the measures of effectiveness established. Rapid growth of air transportation makes capacity a pressing issue, and allows us to describe this purpose in terms of three related goals:

1. To find a precise meaning to the descriptive term "air traffic control system capacity";

2. To find quantitative relationships between the capacity of the air traffic control system and the performance of the air transportation system of which it is a part;

3. To find quantitative relationships between the capacity of the air traffic control system and the characteristics of the elements – equipment, procedures, people, and configuration – of which it is comprised.

This report describes the first year of activity in a program expected to last several years. Much of it has been deliberately exploratory. We have been learning about the historical developments of our air traffic control system, the growing awareness that capacity is an issue, and the present scale of the problem. We have been finding out the aspects of the capacity problem which are of greatest interest to the FAA at present.

Initially, in the first year we expected to complete the following sequence of tasks:

- Identify needs and uses of tools;
- Identify and describe system components and procedures;
- Formulate subsystems relevant to needs and uses of tools;
- Develop detailed block diagrams;
- Determine and develop required terms and measures;
- Evaluate and examine model approaches and computer and data requirements;

- Select and formulate optimal tools;
- Establish data specifications;
- Draw up a detailed 4-year plan; and
- Draft and review a final report.

We have departed from this sequence for a variety of reasons, as indicated in our results. At this point in time, we are continuing a general analysis of the concept of capacity and the way the word is now applied to air transportation problems, and also carrying out work in three problem areas: the analysis of the air traffic control system in terms of goals and functions; a study of the relation of safety to air traffic control system capacity; and a study of the theory of nonstationary queues. These are all important to an ultimate understanding of an air traffic control system capacity, and will form a foundation for a continued study of the capacity problem.

The capacity methodology which will ultimately come out of this study will be used not only to analyze the capacity of systems and subsystems in operation, but also to predict the effects of modifications, substitutions, and new developments in air traffic control. For the definitions, measures, and methods to be useful, they must have predictive value, and must be useful with equipment, procedures, and subsystem organization different from what is presently used. But the various competing concepts and implementations of air traffic control are not totally unrelated: they share the same environment, they operate over the same physical space, they work with the same vehicles, and they are intended to achieve the same goals. In many systems, there is only one, or possibly a small number, of generic sets of functions which could logically lead to the achievement of the goals, although there are many different procedures and physical embodiments which could carry out the function. One advantage of describing a system in terms of its functions is that the number of alternative embodiments is much less than the number of alternate equipment configurations.

A secondary advantage of describing a system in terms of its functions is generality. On the one hand, a valid analysis of the function is simultaneously an analysis of certain aspects of any embodiment of it. On the other hand, a vivid characterization of the function may suggest a variety of alternatives for executing it other than those traditionally used.

Another reason for seeking descriptions in terms of functions is their relation to performance criteria. It will be shown that purposes and functions are complementary in the sense that the purpose of a subsystem is commonly a paraphrase of a function of the system of which it is a part. But it is also true that criteria of performance and measures of the overall quality of performance can also be phrased in terms of goals and purposes, the degree to which they are achieved, and the cost of achieving them. This provides a bridge between the system description and a description of overall quality of performance. It will be shown later that an understanding of the overall quality of performance of the air transportation system is a necessary ingredient in the definition of air traffic control system capacity.

As the various steps in the creation of a description of air traffic control in terms of its functions are carried out, we shall examine each function to determine whether its implementation may constrain capacity. A constraint may arise because of the intrinsic character of the function, or because of incidental properties of its implementation. For example, as long as final approach and take-off separation are motivated by the necessity for avoiding simultaneous runway occupancy, a terminal area separation function will constrain terminal area capacity, regardless of the means chosen to implement it and the technical perfection with which it is carried out. It seems quite plausible to consider terminal area control schemes in which the actual spacing between aircraft is only very slightly larger than the minimum required to avoid simultaneous runway occupancy. An appropriate quantitative tool to study the resulting constraints on flow is queuing theory, especially the theory of non-time-invariant queues.

En route airspace, unlike a runway, is roomy enough for vastly more operations than are accommodated by present usage. Separation standards impose a real limitation on flow here as well, but the standard of safe separation cannot easily be referred to a simple criterion such as interdiction of simultaneous runway occupancy. It must be referred to the probability of mid-air collisions under circumstances where, although they are extremely rare, their probability is not negligible. The margins in space and time required to assure that collisions are sufficiently improbable are large, and depend sensitively on the tenuous probability distribution of measurement and performance factors far from their control values. Before a queuing or other flow model can be satisfactorily applied, we need a quantitative understanding of the relations between collision probability, separation standards, and other operating parameters.

As a third example, we may note that air traffic control, as presently implemented, is labor-intensive; that is, a large proportion of the total cost of air traffic control is represented by salaries, with the salaries of air traffic controllers and their immediate support making up the largest part. Under these circumstances, good management inevitably requires that the human components of the system be operated at near their maximum capacity. Otherwise, substantial economies in system operation could be effected by personnel reductions. We can predict, therefore, that controller workload will put an actual constraint on capacity, and that an analysis of the capacity of the present air traffic control system cannot be complete without a quantitative understanding of the relation between the amount of traffic and the controller workload. We have not, however, undertaken to study controller workload.

During the early months of the study we actually worked on functional descriptions, time-varying queues, and safety simultaneously rather than in logical sequences. We expect that further work on functional descriptions will show where further quantitative tools are needed, what characteristics they should have, and how they fit into a balanced study of air traffic control systems capacity.

The relevance of safety to air traffic control system capacity has never been doubted, but neither has their relation been clearly enunciated. In very simple terms, the immediate effects of puttinp more aircraft in the air without making any other compensating changes or adaptations is to increase the risk of collision in the air or on the landing strip. However, while the size, number, and speed of aircraft have all increased dramatically in the last 15 years, the rate at which accidents take place has gradually and slightly decreased. Safety has been stabilized at an acceptable level by introducing technical improvements and by using other service degradations like delay to reduce risk. Any change in equipment or operation which reduces risk could also be interpreted to increase capacity, for we could eat up the new safety margin by adding more traffic to the system. If we ever expect to attach a quantitative measure to capacity, we must have a quantitative understanding of the trade between hazard and other operating parameters such as separation which ultimately figure in a capacity determination.

In most trade-off analyses, we attempt to optimize some function of an assortment of costs and values subject to some constraints. It is common to find conflicts where increasing one value degrades another, so that a compromise must be achieved between the two. In the case of an air traffic control system, the trade between amount of traffic and certain other system degradations such as delay is not direct, but indirect through the action which each has upon risk. When the relation of each with risk is fully understood, it may be possible to eliminate risk as a variable by treating it as a rigid constraint. Even though this may become possible, it may still be preferable to treat it as an explicit parameter. At present, however, our understanding is so limited that neither alternative is possible.

The importance of time-varying queueing theory to the study of air traffic control system capacity is somewhat more transparent. We shall show that capacity is related to system overload which makes itself evident by some performance degradation. Many of the immediate causes of system overload in air transportation are transient or ephemeral phenomena and the behavior of the system cannot be understood in terms of long-term averages. Actual waiting time and delays are neither much larger than nor much smaller than the times required by the air traffic control system to take corrective actions, to institute flow controls, and the like. Actual demands vary in diurnal cycle, and certain aspects of system performance can change rapidly with changes in weather or visibility. Thus the relation between the demand on a subsystem and its performance capability may change significantly in a period of time comparable to the time delay in some significant control loops, the delay time in a queue, and the travel time under a flow control discipline. For these reasons we would expect the dynamics of the interaction of a queue with varying demands and service parameters to be significant in explaining system behavior, and that a characterization of performance in terms of long-time averages and steady-states would be inadequate.

2.3 RESULTS

We have made progress in identifying and describing system components and procedures, formulating subsystems relevant to needs and users of tools, and developing block diagrams. Available descriptions of the air traffic control system and its subsystems are not germane to the issue of capacity. Most of them describe physical layout, geometric configurations, procedures, and engineering specifications of equipment, with very little explanation of why the air traffic control system is put together the way it is.

We have made some progress in describing the various elements of air traffic control in terms of function, needs, and means rather than in terms of equipment or physical configuration.

As anticipated, attempts to describe the functions of the air traffic control system have stimulated the formation of concepts and the definition of terms. We have devoted special attention to the concept of capacity, and have found it desirable to give extra attention to the concept of safety and its quantitative measurement.

The tasks concerned with evaluating and examining modelling approaches and computer and data requirements, with selecting and formulating optimal tools, and with establishing data specifications have not been carried out in a form corresponding to the descriptions in our proposal of a year ago. We have, however, initiated two other lines of investigation: one, into the role of safety in air traffic control system capacity methodology, and the other, a review of some aspects of the theory of time-varying queues.

Among all of the system performance criteria related to capacity, safety has been the most difficult with which we dealt, and it is probably the most important. Wherever the influence of safety is felt at all in decision-making, it has priority over other considerations. A very large proportion of the decisions made in air traffic control are justified by appealing to safety as a motivation. Yet the actual number of fatal accidents is so small that it is almost impossible to base nontrivial conclusions validly on accident statistics. Thus, there is a need both for theoretical models of safety and for indirect quantitative measurement techniques.

It is well known that standard models from queueing theory can describe many phenomena in air traffic. In some applications of queueing theory (for example, the study of congestion in telephone switching centers), service and waiting times are short in comparison with the time required for a substantial change in environment, operating conditions, or demand. In air transportation, the situation is quite different. Substantial changes in demand and in capacity to render service often take place in a fraction of an hour, invalidating a queueing model which assumes steady-state conditions. A common response to this challenge has been to use simulation. However, there is a considerable body of mathematics available to deal with time-varying queue systems, but almost none of this has been adapted to air traffic control system problems. We are exploring the usefulness of time-varying queue analysis to air traffic control capacity problems by formulating and analytically solving some illustrative problems.

3. ESTABLISHMENT OF TERMS AND MEASURES

3.1 INTRODUCTION

To establish a comprehensive, well-defined set of terms and measures with which to describe and evaluate the air traffic control system, we must first analyze general concepts and render them in precise terms. This chapter deals with the concepts of capacity, demand, and delay which prove to be inseparable. Other concepts will be examined in later chapters of this report, particularly safety and risk and related notions.

The process of defining terms and measures is iterative. Relevance, intelligibility, and measurability are the criteria for the choice of a first tentative definition. To improve on these definitions, it is necessary to form a precise conception of the mechanism underlying the system under consideration, e.g., the ATC component, the air terminal, and the like. This step is often called "setting up a model." Next, implications regarding the quantities introduced under the terms and measures are examined in the light of this conceived mechanism: by "operating the model." After enough examinations of this sort, the degree of adequacy of these terms and measures to express organic features of the situation become better understood. Moreover, certain *other factors* may become apparent, also requiring precise definitions, but which may have been missed in the initial formulation of terms and measures. With the improved list, more relevant models can be set up and operated. This interplay of terms and measures with models and measurements – this process of cyclical refinement – is common to all developments of science and its technological applications.

The general concepts of capacity, demand, and delay correspond, respectively, to (1) how much an element of the air transportation system can handle, (2) how much it is requested or desired to handle, and (3) the disadvantage – in terms of time lost – that this handling may incur. These concepts, as we have found, cannot be entirely separated and developed in isolated compartments.

The issues are brought to a focus in the recognition that the *problem of air transportation capacity and demand* is a problem of the *allocation of a scarce commodity*. Therefore, the establishment of definitions (terms and measures) must recognize the problem itself: how much of the commodity is available, how much is wanted and by whom, and how much degradation of the commodity (delay, etc.) is acceptable.

Preceding page blank

3.2 THE CONCEPT OF CAPACITY

In connection with air transportation and air traffic control we can consider three different meanings of the term "capacity."

First, capacity can be considered as a static *"holding"* or "container" capacity. This may apply to real entities, such as taxiways, airport terminal gates, or aircraft holding areas, as well as to abstract entities, such as information lists in a mechanical data processing system, or the span of control of a single controller. In some cases the level of such a static capacity will be determined solely by the available "space" and the nominal or physical "dimensions" of each entity. In other cases, the capacity will be a function, too, of the extent and type of interaction between occupants of the space which, in turn, may be a function of external parameters. Entities such as holding areas have capacities which depend both on the geometry of flight paths and stack management rules as well as on the acceptable level of space occupancy degradation or crowding. This is analogous, for example, to a bus, the holding capacity of which may be a function of the number of stops it makes and the related internal movement and the level of congestion that passengers will tolerate.

Second, capacity can describe a *rate*. This is a time analogue corresponding to the holding or container capacity. Whereas holding capacity is defined by the match of available space and the physical or nominal dimensions of entities, rate capacity involves events which have a time dimension and a certain available time in which so many events can be contained. Most of the subsystems and components of the air traffic control system – both in its real physical flow embodiment or seen as an information handling system – have rate capacities: the numbers of events that can be accommodated in a certain time period. Of course the mix and specification of the events, assuming that they are not all identical, will be important determinants of the rate at which the subsystem or component permits events to occur. Whether the subsystem is a controller processing handoffs, or a data processing system handling flight plans, or a glide path-runway-taxiway subsystem accepting arrivals and departures, the same concept is applicable.

Third, the term "capacity," corresponding more directly to one of its everyday uses, can be used to reflect the overall capability of a system or subsystem to perform a given quantity of a particular task at a certain "quality." In a system containing both elements limited by a holding capacity and elements limited by a rate capacity, which is subject as a whole both to considerable demand fluctuations and variations in the parameters which define the capacities of the subsystems, there will be a (upper) level of demand of a given distribution over a significant time period, say, covering at least one major demand cycle, at which the overall quality of the performance of the task reaches some (lower) feasible level. This third type of capacity is, it seems, normally referred to when the expression "capacity of the ATC system" is used.

Capacity always means a bound at or near which some kind of overload occurs. When the value of a parameter is below its capacity, the situation is "normal." When the value is greater than its capacity, the situation is "bad." When we say that "the capacity of this bottle is one quart," we understandthat an attempt to pour more than one quart of fluid into the bottle will result in spillage. In such a simple case, we need not state explicitly the consequences of exceeding the capacity. However, in more complicated situations, we cannot rely on intuition to define the overload. The definition of capacity is incomplete without the specification of the form of overload; i.e., what goes wrong, and how, when the capacity is exceeded? The specification of the consequences of overload is particularly important when talking about capacity in the third form mentioned above, that which reflects the overall ability of a system or subsystem to perform a given quantity of a particular task at a certain "quality." We shall see later that the quality of air transportation services has many dimensions, and much of the problem of defining capacity results from the trading-off among various criteria of service quality.

3.3 CAPACITY OF AN AIR TERMINAL

As a concrete example, we shall discuss the capacity of a single air terminal.[1] From the point of view of aircraft handling, an air terminal is called on to receive *incoming* aircraft and serve *outgoing* aircraft. If very few of either request such service, there is no problem of capacity: this problem arises only when the traffic increases enough to tax the terminal's facilities. With most air terminals, the ability to handle outgoing aircraft is diminished, while it is allowing many aircraft to land; the number falls from a *maximum,* when no aircraft are coming in, to *zero,* when the terminal is devoting itself solely to the landing of a heavy flow of incoming aircraft. A similar statement can be made for landing aircraft. The situation may be expressed precisely in terms of Model A described on the next page.

3.3.1 Model A. Assume that the air terminal is being operated during a certain period under constant conditions of such a high rate of landing and take-off demands that its facilities are always fully used. If the average number of aircraft admitted to land in unit time is u, and of those taking off is v, the total mean number handled (events) per unit time is represented by u + v. In one sense, this represents the *capacity* of the air terminal, but it is too much to assume that this *total handling rate* u + v will be independent of the *mix* or ratio u/v. Thus the maximum landing rate u_0 (value of u when v = 0) and maximum take-off rate v_0 (value of v when u = 0) may be different from one another and still different from the intermediate measure u + v (when neither u nor v is zero). On the basis of this, one is led to the following definition:

> *The instantaneous capacity for any given ratio u/v of allowed landings to take-offs is the maximum possible number of events u + v of landings and take-offs per unit time under constant saturation demands by both classes of aircraft.*

Naturally, when different *types* of aircraft are served by the same airport; e.g., V/STOL, piston and turboprop, subsonic and supersonic (SST) jets, the capacity in the above sense will reflect the ratios of aircraft of different types. Without going further into such matters, we submit that the above version of capacity is clear, measurable, and relevant to the study of different modes of using an air terminal.

One logical reservation must be made in applying this definition, as well as those given below, and indeed wherever the concept of "maximum" possible number of events, etc., is used, since a practical maximum always implies certain practical constraints, and these may not be easily quantifiable. Thus a controller may accept a high rate for a brief period (e.g., 15 minutes) which he will refuse over a longer period. This caveat will recur later.

The precondition in Model A is that the facilities of the terminal be *fully used.* This condition must be understood in a relative rather than in an absolute sense: fully used under certain constraints of allowable delay, and so forth.

The simple model just studied is useful up to a certain point. In actual practice, the demands for take-off and for landing are never at a saturation rate during every hour of the day. During the night and early morning there is no demand, or so little that the issue of capacity does not arise; during the morning and afternoon "rush-hour" traffic on weekdays, there is often, even under favorable weather conditions, such a volume of demand that waiting lines on the ground or stacking queues in the air may form. This leads to delays; but as long as they are not too great, all the traffic can be handled if the aircraft wait until the peak passes. The situation described has (during weekdays at least) a periodically repetitive character having a 24-hour period. There may also be a 7-day weekly period. These are familiar facts in the study of automobile traffic in highways, in which the problem of capacity is also a serious one. Model B described below takes the 24-hour diurnal periodicity into account, and makes the assumption (sometimes realistically and sometimes not) that there is no restriction on the number of aircraft that can gather in ground queues or in air stacks.

3.3.2 Model B. Suppose that the rate of arrival of aircraft intending to land at the terminal is a known function $a(t)$ of the time of day t, having a 24-hour period $[a(t + 24) = a(t)]$. Suppose similarly that the rate of entering the take-off waiting line is also a known function $b(t)$ with the same 24-hour period. Further, suppose that the air terminal policy fixes on a particular ratio u/v of landings to take-offs (during high demand periods). Finally, suppose that there is no limit to the number of aircraft allowed in either the air or the ground queue.

The simplest analogy of the flow of material fluids through reservoirs provided with pipes and orifices indicates that there are two possibilities: (1) either the accumulation of aircraft in the queues (fluids in the reservoirs) during the times of high traffic (most rapid inflow) is able to pass out of the system during low traffic periods, so that by the time of lowest input they have all left; or else (2) this is not possible: The 24-hour accumulation in at least one reservoir augments indefinitely. At or just before this stage of saturation, there is a total 24-hour number $U + V$ of aircraft that are passed through the terminal: U landing (at the mean rate $u = U/24$) and V taking off (at the mean rate $v = V/24$). Thus the strict analogy with the flow of fluids would suggest that the total rate u + v is the same as the *instantaneous capacity* defined previously: it is achieved, however, at the expense of more delay of the aircraft that intend to land or to take off during peak periods.

Figure 3-1 illustrates the situation graphically by plotting the number of aircraft arriving for service (landing or take-off) per unit time vertically in the two cases: that of constant saturation rate (the horizontal line A) and in the case of a 24-hour periodic rate (the curve B). If the terminal can just handle the air traffic in the latter case, the total area under the latter curve must equal that under the straight line. This total area represents the total number of aircraft handled (events), and when divided by the base (24 hours) represents the mean rate of handling events. This is the ordinate to the straight line and the *mean* ordinate up to the curve. Note that in the fluid analogy, what is here shown as an area would be represented as a volume of fluid.

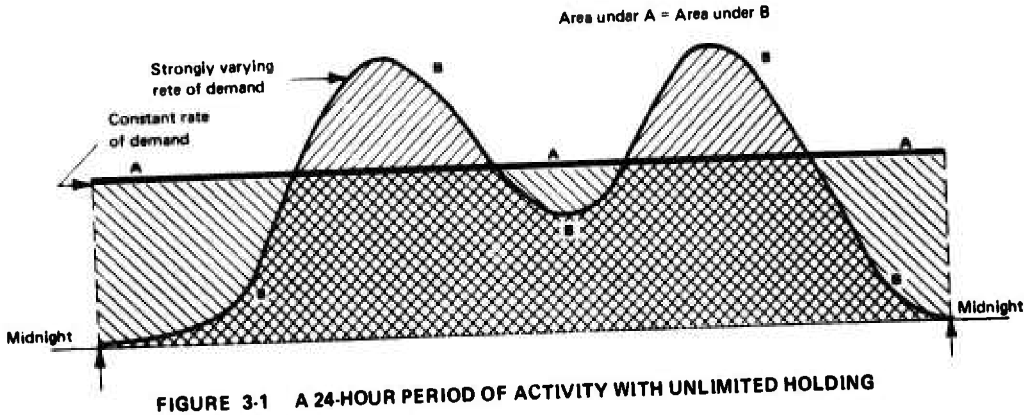


FIGURE 3-1 A 24-HOUR PERIOD OF ACTIVITY WITH UNLIMITED HOLDING

Thus with the situation of Model B, the capacity is still defined as before, but with the difference that a 24-hour average must be considered.

Here, again the logical reservations are in order regarding the use of the concept of "maximum" rate. The situation is further complicated by its application at the peak demands: a rate may be accepted as a momentary excursion if it is perceived as a chance fluctuation, which might be refused if it persisted – pilots and controllers will cut a few corners to squeeze in a few extra operations during a short peak, rather than cause missed approaches and wave-offs which increase queues and traffic loads without achieving a landing.

The situation assumed in Model B, in which no restrictions are placed on the number of aircraft in either the ground waiting line or the air stack, is only realistic in the case of a slight peaking; i.e., when curve B of Figure 3-1 is quite close to the horizontal line A, since then the waiting line or stack will never be unreasonably long. But such a situation, combined with a 24-hour saturation, is not apt to arise. Therefore we pass to Model C.

3.3.3 Model C. Everything in Model C is the same as in Model B, except that there is a limit (l) to the number of aircraft allowed in the ground queue and also a limit (m) to the number allowed in the air stack. Any aircraft that seek service are either rerouted or held at their place or origin, if to admit them would cause longer queues than (l) or (m).

Continuing with the fluid flow analogy, we may think of the vessels into which the fluids representing the aircraft seeking to land and those wishing to take off flow as being open and of limited volume: when the fluids are poured into them faster than they can pass out through the openings representing their accomodation by landing strips, they simply spill over the top (are diverted).

Figure 3-2 represents this situation, with conventions similar to those of Figure 3-1. The horizontal line A again represents, by its height, the maximum rate of throughput (events that the terminal can accommodate; i.e., its instantaneous capacity at the given ratio u/v). Curve D (dotted) is the "demand curve," i.e., its ordinate at any time t is the rate at which aircraft (taking off or landing) would wish to be served. If the area under D were equal to that under A, then D would be curve B of Figure 3-1; but we are thinking of it as possibly having a greater area. It is the graph of a(t) + b(t) against t. The actual rate of arrival of aircraft that can be accepted by the system without exceeding stacks and waiting line limits is given by the curve C, obtained by removing parts of the area under D, as shown. After the number in the two queues (area developing above A and below D) reaches its allowed limits, aircraft are admitted only at the rate which can be handled without further increasing the waiting queues; therefore, we cut D down to A. When this is no longer necessary, C is allowed to run along D again.

If a vertical line is drawn to cut the horizontal axis at a point corresponding to time t, the length of the air and ground queues at that time is found by taking the total shaded area above A and under D to the left of this line, and subtracting from it any area to the right of the shaded area and to the left of the line which is below A and above D, when the latter is less than the former; when the latter is greater than the former, it is zero. By this process the ordinates (length of queues) in Figure 3-3 are obtained. If a second peak occurs before the queue from the first peak is dissipated, an obvious modification of this process is required.

Simple as this model is, it reveals an essential fact facing any attempt to attach an all-purpose single measure of "capacity" to an air terminal. The fact:

> *The number of events that can be handled in a 24-hour period may depend strongly on the shape of the demand curve D.*

To show this effect graphically, suppose that the demand curve D of Figure 3-2 were replaced by the curve C constructed in that figure. In other words, suppose that the applications for landing or take-off that had to be refused with the original D were prevented from existing, all others being as before. Since the resulting curve C represents a rate of events that can be accommodated by the terminal, the area under it cannot exceed that under the horizontal line A. Otherwise there would be a contradiction with the construction of curve B of Figure 3-1, the maximum rate of arrival that – even under the less stringent conditions of Model B – could be accommodated: and the area under B equals that under A. To emphasize the dependence of the number of events with which the terminal can deal in a 24-hour period upon the shape of the demand curve D, consider the exaggerated case in which D does not rise above zero except during two hours of the day (e.g., 8-9 a.m. and 5-6 p.m.), but has the same area below it as A, and with very strict waiting line limits: the area representing the event capacity could fall to little over one-twelfth (2/24) that of the uniform arrival rate A.

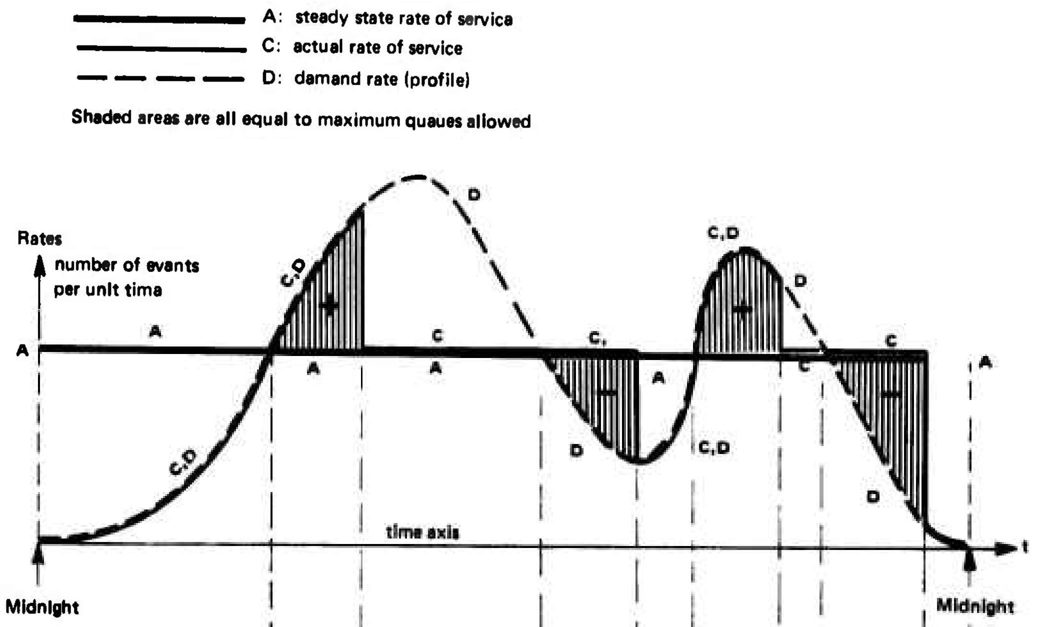


FIGURE 3-2 A 24-HOUR PERIOD OF ACTIVITY WITH LIMITED HOLDING

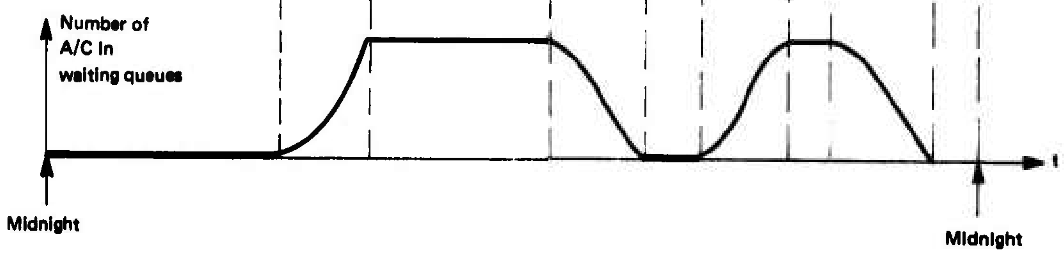


FIGURE 3-3 THE WAXING AND WANING QUEUES OF FIGURE 3-2

A second fact has to be noted in this connection: Suppose that a demand curve D, while strongly time-dependent (and having a 24-hour period) never rises so high that aircraft cannot be accommodated (possibly after some time in queues). The corresponding curve C of Figure 3-2 could then be taken as this same D without alteration, but the area under it would not necessarily give a valid representation of the airport's capacity, since it might merely be the result of *under-use* of the latter.

Taking all these facts together, we return to the original definition of capacity – given after Model A – as the rate of event occurrence (activity) of the airport during a period of saturation. Provided that the relative nature of the condition of "saturation" or "maximum use" is kept in mind, as emphasized before, this is a valuable *first step* in the formulation of the definition of the "capacity" of an air terminal.

The definition and the background discussion have been confined to the case of a single air terminal. Obviously they have to be extended in two directions: (1) to the *component parts* of an air terminal, such as the runways, the ATC system at the terminal, and many of the other factors which, operating in unison, generate the capacity of the terminal; and (2) to cooperating sets of terminals, such as those in the Golden Triangle, the Chicago-to-NE Area, and so forth. Again the notion of full practicable use and the maximum rate of handling – either instantaneously (steady state with unchanging demands), or averaged over a 24-hour period – represent the key to the term and its measure.

3.3.4 Stochastic Models. Up to this point, the concept of continuous *flow* (steady or periodic) has served as the basis of the models and related definitions of capacity. We should now take a further step toward realism, and recognize that the arrival of aircraft at a point where they seek service is not only unlike a continuous flow in being "lumpy," but represents a sequence of events having a considerable element of *random*. Only by thinking in terms of *averages* (strictly: *expected values*, in the sense of probability) is the semblance of a deterministic flow of a fluid restored – and the above definitions of capacity meaningfully given.

Returning to the case of a single air terminal, we must realize that even under fixed weather conditions (always VFR or IFR), we cannot say just when aircraft will arrive for landing. Take-offs may commence with more regularity, but their servicing by the runways under congested conditions, and taking turns with random arrivals, soon communicates the element of random to these operations. Furthermore, any unpredictable events occurring in any part of the air transportation system, of which the air terminal is a part, will cause changes from its average states. Wind, instrument variability, navigational uncertainties, and the exercise of the pilot's option of choice among not fully specified flight plans all contribute to randomizing the flow.

The problem of describing these circumstances, with their mixture of regularity and random, is too complicated ever to be solved accurately and completely. Simplified models have to be used, and these must be able to handle the predictable (often time-dependent) features in combination with probabilistic ones. The results are the various stochastic processes which are discussed in Chapter 6 of this report (references to the literature are given there). Only after enough such models have been set up and analyzed can a firm basis for further refinements of the basic terms and measures be established.

Without awaiting the results of such a technical examination of the random factors, however, *the mere recognition that they exist* allows us to draw certain qualitative conclusions regarding the concept of capacity – and later – those of demand and delay.

First, if the "rate of flow," which was used as a building stone in our earlier definition, is recognized only as an expected value of a fluctuating quantity, then attention is automatically directed to the *amount of dispersion* of the latter quantity about its mean: is it large or small, predictable or unpredictable, and how does it behave under changing conditions? Precise answers can only be found on the basis of probabilistic work, such as that in Chapter 6, or of lengthy and systematic observations going even beyond those tabulated in Reference 2. Nevertheless, some quantity, such as the *standard deviation* of the aircraft arrivals, or the like, must find a place among the basic terms and measures, since it is related to both capacity and demand.

Next, recognition of this random element forces one to face the question of the *stability* of the state of affairs underlying the simplified flow concepts upon which our earlier definition of capacity – as a maximum fully-utilizing rate – was based. An actual step-by-step examination of methods of maximizing the utilization of a system of runways at an air terminal has shown that *with an increase in efficiency there is an inevitable increase in instability.* This is almost a general principle of operations research, and shows the practical fallacies that may face suboptimization. Reference 1 shows how the introduction of general aviation units into a system with a nearly saturated terminal may cause delays in the whole schedule quite out of proportion to their numbers. Obviously, *stability has to find a place among the basic terms and measures,* but only after further observational and mathematical study.

Finally, these considerations make it necessary to keep the requirement of stability as a constraint in defining "capacity" as maximum utilization. Actually, most ATC operators tend to keep stacking spaces in reserve – not to fill them to their physically maximum possible extent – to act as a buffer against some fluctuations and avoid instability. These are the factors reflected in the idea of a "*peak capacity.*"

3.3.5 The Units of Capacity. It is evident that in defining the capacity of an airport – or of systems composed by it or composing it – in terms of *events* (landings and take-offs or aircraft), one other useful possibility should be noted; viz., the number of *units transported* by the aircraft, e.g., *people,* tons of *merchandise,* and the like. Then, for example, the "capacity" as a number of transported units could be increased without changing the "capacity" in the sense of number of events -- by using more efficient types of aircraft. This would be of only indirect interest to the ATC problem in its narrowest sense, i.e., as weight of work of a control tower, which is interested directly in the number of events.

In conclusion, we have defined *capacity* as the practical maximum 24-hour mean of units handled – events or transported objects: that is, under agreed-on limits of liability to instability, delay (see below) and risk (see Chapter 5). Related to this quantity are *peak capacity* and *dispersion* – quantities implied above but to be defined in terms of the probabilistic analysis in Chapter 6.

3.4 DEMAND ON AN AIR TERMINAL

The demand in the sense of the rate of applications for landing [a(t)] and take-off [b(t)] made on an air terminal has been plotted as the *curve* D of Figure 3-2 in Section 3.3. The shape of this curve represents the hour-to-hour rate at which use of the air terminal is desired by the public, and it can be ascertained by statistics. But the steady-state rate of service A cannot be substantially reduced without impairing the service rendered to the public.

In general planr.ing, however, the mean demand rate over a 24-hour period may be a useful single numerical characterization of the degree to which use of an air terminal may be sought. Since the area under demand curve D of Figure 3-2 represents the total demand for the air terminal's services by aircraft arriving or departing, the *24-hour mean demand rate* is this area divided by the 24-hour base.

Thus the "demand" as a curve – which will be called the *demand profile* – and the demand as an average number (area under this curve divided by 24 hours), the *mean* (daily) *demand*, are both useful. The former applies to the evaluation of the burden of operation facing the airport, its ATC system, and so forth; it is relevant to the planning of optimum use. The latter, the mean daily demand, is relevant to any forecasting of the general facilities that should be installed for handling aircraft and similar overall planning.

The same remarks are in order as at the close of Section 3.3: as a *unit* in these definitions we might also use the person or weight of goods transported; to be exact, the definition of demand profile or mean *must have its unit specified.*

Both the demand profile and the mean demand may require refinement; there are as many demand profiles as there are important classes of aircraft that may wish to use the terminal – air carriers, general aviation, air taxi, military aircraft, and the like – and, correspondingly, many mean demand numbers. An intensely important practical problem is the manipulation of the various profiles (by general regulations and agreements, and so forth) so as to respond to the total set of mean daily demands, under the constraints of stability, delay, safety, and similar factors. The background discussion of this matter is developed in Reference 1.

The flow picture (Models A, B, and C) serves as a basis for the *definition* of demand as we have given it. The more precise methodology and analytical tools of later chapters are needed to examine optimum methods of responding to this demand. Finally, each step of the process requires the statistical observations of airports.

Up to this point, "demand" has been interpreted in the narrowly focused sense of what is required of a particular element (e.g., the air terminal) of the air transportation system. It would be shortsighted to omit a broader viewpoint: the air transportation system is itself just a part of the full national (and international) system of transportation of people and goods. Also air transportation is a *scarce and desirable* commodity. If it were of unlimited availability and as cheap as any method of transportation, the "demand" would be extremely great; no other method of transportation would be used in most non-pleasure operations. In this sense, *the "demand" is unlimited* – i.e., always exceeds the capacity of any foreseeable air transportation system. The actual demand limitations result from the cost, the limited capacity of air terminals, and the *inaccessibility of air terminals* to so much of the country. Attempts to increase the number, size, and accessibility of terminals would come up against civic constraints. Even if these did not exist, saturation of airspace would become a constraint. In the light of these and similar observations, we may recast the concept of *overall demand* as the demand for air transportation that would actually occur within the cost structure and restricted availability under civic and safety constraints.

3.5 DELAY AT AN AIR TERMINAL

In every system of transportation, two quantities, in an essential way must be considered as characterizations of its effectiveness: (1) the *bulk transportation rate*, or number of units transported from the point of origin to destination (i.e., which cross any fixed plane separating them); and (2) the *speed of the transportation*, which is the mean distance each unit travels per unit time in moving from its point of origin to its destination, including time spent in waiting lines. The first may be large without the second necessarily being large (as when there are many slowly moving units in the "pipeline") and vice versa (a very few rapidly moving units). In train and ship transport, the bulk rate has usually been higher than in air transport, while the opposite is true for speeds. The concept of delay corresponds to an increase in time taken over what would normally be expected, due to an untoward fall in speed. It is a substandard quality of service.

Suppose that an aircraft reaching a standard distance (e.g., 50 miles) from an air terminal could – if there were no other aircraft using the terminal at that time, and if all other conditions (weather and equipment) were favorable -- make a landing after a time T (e.g., 10 minutes). This might be called the "standard minimum" time, and could not reasonably be regarded as a "delay." But suppose that under less favorable conditions, as when it is necessary to await other aircraft to land or take off from the field, or when ATC equipment is saturated, it may take a longer time T'. Then the difference T' - T can be definited as the *delay in landing*.

A similar definition is given for delay in take-off: the *actual time* taken to join and remain in the take-off waiting line, and then to get airborne and fly to the standard distance from the air terminal, *minus the minimum of this time* under perfectly favorable conditions.

Both delays in landing and take-off are evidently numerical measures of a type of degradation of service. They can be found observationally by gathering suitable statistics. On the other hand, to predict their values in projected situations, for the sake of aiding in planning the introduction of material improvements and of optimizing the utilization, a clear quantitative conception of the mechanism of the system – i.e., a "model" – is the necessary starting point.

Models A, B, and C, introduced earlier, could be used to give a first approximation to the prediction of delays under various conditions, using the flow analogy, but only after supplementary assumptions are made regarding the diminished rate of advance of an aircraft as the utilization of the system increases (i.e., when it is operating at full capacity). This would change the picture from the one of flowing liquids to one, rather, of flowing gases, suffering compression. The model would become artificial and not a reliable simulation of reality.

For the reasons set forth in the last paragraph, as well as for those adduced toward the close of Section 3.3, the flow models must be replaced by others which are closer to reality. As stated before, aircraft do not enter and pass through the system (the air terminal) in the manner of a fluid. They have a *strongly random* element in their arrivals and in their processing, and it is only in their highly conventionalized *averages* that they present a picture of a flow.

Such improved models, taking into account random events and dealing with the probabilities, will be considered in detail in Chapter 6. As stated earlier, the *average* values will not only be handled by the techniques of stochastic processes, giving a more solid basis for the definitions of capacity, demand, and delay, but certain other quantities will also enter, representing the fluctuations of these variables away from their averages. Of course, the strongly time-dependent effects will also be taken into account.

Then it will be possible to calculate the *delay* in the sense of the *expected or mean value* $\bar{T}' - \bar{T}$ of the time-excess quantity T' - T introduced in our first definition. It will also become possible to predict its behavior under various actual or hypothesized operating conditions. Not only will this mean-time excess through the system become numerically computable, but its dispersion (e.g., standard deviation) will become an output of the mathematical methods (see Chapter 6). Finally, a basis will be obtained for comparing the results of statistical observation at the air terminals, and so forth, with these stated outputs of our analytical tools.

With reference to the opening paragraph of this section – the *bulk* transportation rate versus the *speed* of transportation – a corresponding duality exists in the two characterizations of air terminal performance, *capacity* and *delay* (or its opposite – speed). Moreover, at a high rate of demand, by increasing capacity any response tends to increase delay as well. For example, if unlimited queues were made physically possible, so that Model C of Section 3.3 could always be replaced by Model B, the capacity would be increased; but for many aircraft this would mean long delays waiting in queues. After a certain point, even if the aircraft had the requisite endurance, it would become quicker to "go by train," and the demand would fall off. In making such a choice, prospective passengers logically would not only have to compare the *expected* times taken by the two methods of travel, but would also have to take into account the *dispersions* in these times; hence, a point of practical importance of defining "delay" in terms of probabilistic rather than deterministic models.

Up to this point, delay has been discussed as an attribute of the response of an element in the air transportation system when handling various demand profiles. One could be serious'y misled if he were to overlook another factor in the delay picture; viz., the overall loss of time incurred by passengers, and indeed by "potential passengers," who are unable to fly when they want to because of airport congestion. To illustrate, suppose that it is decided that delays at a particular airport are due to rush-hour peak periods, and that (to oversimplify) to provide many flights between 8 a.m. and 9 a.m. and 5 p.m. and 6 p.m. leads to delays in waiting queues; and hence a decision to spread the same flights evenly between 7 a.m. and 11 a.m. and 3 p.m. and 7 p.m. is made. Even if each flight on the new schedule experiences zero delay, the people for whom the schedules exist may now experience even more loss of time on the average than they did with the original schedule: those who cannot find a place on a flight may have to take an earlier one, and waste an hour before the office at which they wish to do business is open; or a later one, and lose a valuable hour for business; or, finally, they may have to travel the night before to be able to avoid undesirable times of arrival. Similar losses could be incurred by the need of accepting inconvenient schedules at the close of the day. Evidently, therefore, *the problem of air transportation delays must be viewed more broadly than in terms of slowed take-off-to-landing times.* Such broader considerations could easily lead to a decision to spend money for additional ATC and runway facilities, even when the point-to-point delays could be avoided – at the expense of inconvenient schedules.

3.6 CONCLUSION

The program set forth in the Introduction has been carried through the first cycle, based on the simple model of continuous flow:

> *Capacity* has been defined as the rate of accepting a maximum, fully acceptable rate of a steady flow of aircraft seeking service, or of its average in case it varies with a daily period – all under the numerous practical constraints. It has been recognized as a quantitative characterization of the way in which the terminal responds to any given schedule of demands.
>
> *Demand* has been formulated both as a schedule – the demand profile – and as a number – the mean daily demand.

Delay has been defined as the increased time of service under the given conditions as compared with ideal conditions. Just as much as capacity, it is a numerical characterization of the terminal's response to a schedule of requests (the demand profile) – and under the same practical constraints.

The discussion has indicated the necessity of replacing the deterministic flow model by a more realistic stochastic one, thus recognizing the random nature of the problems facing the terminal. While this will be treated later, many qualitative factors are put in evidence by the concept of *random*: the magnitude of the disposition about the mean; the stability of the system of flights; peak capacity, demands, and delays. The question of safety, to be treated later, is also connected with the possible effects of random.

The terms, measures, and related concepts of this chapter have been developed in connection with the relatively simple case of a single air terminal, and one having a single runway and single line for take-off aircraft. The purpose of this restriction is to bring out the concepts in all clarity and concreteness. A corresponding simplification will underlie much of the stochastic waiting line work of Chapter 6. We have found that the same methods and concepts apply, with obvious extensions, to more complex terminals, such as those with several runways and corresponding disciplines. With the flow model, one introduces a few more connecting pipes, while with the stochastic waiting line model, more transition possibilities have to be recognized: whereas, the practical computation grows in complexity, the concepts underlying the terms and measures remain unchanged.

In contrast to the above situation, difficulties of an essentially different order attend the extension of the terms and measures and underlying methodology to more extensive air transportation systems, such as the Northeastern region or the Golden Triangle, composed of many distinct terminals, the air spaces between them, and the full air traffic control system regulating them all. These cannot be understood merely by the study of their separate pieces: there is a complex interaction among the latter, and a *system point of view* has to be developed. It is easy to illustrate the issues involved by a simple example: Clearly if a queue of stacked aircraft wishing to land at LaGuardia exceeds a permissible length, aircraft from other points destined for LaGuardia may be held on the ground, and later create take-off queues, which might not otherwise have existed. Another possibility – particularly if visibility at LaGuardia is slowing landings there – is for aircraft to be diverted, e.g., to Newark. Thus the conditions at one terminal may cause ground queues at others, and increases in landing demands at still others.

While these facts are fully known and understood qualitatively, it is in their quantitative implications that they are anything but simple and trivial. The various models of flow representation, set forth early in this chapter, and which ignore the random elements in demand and capacity, become a more and more tenuous basis for prediction: If we represent the traffic in several air terminals by the model using interconnecting pipes, the resulting behavior will be strongly dependent on the "pipe discipline" -- what decisions are made for switching the overflow of a particular reservoir through pipes connecting to others (i.e., rerouting in case of stacks and delays). But since these occurrences contain a random element which increases rapidly in importance the more subsystems (air terminals) are aggregated into the regional system, it is less and less possible to represent the behavior of the latter by the deterministic (flow) model, in the measure that its complexity increases.

If the more realistic model, which includes the random effects, is used, it is necessary to examine mathematically the consequences of compounding the single air terminal cases examined later in Chapter 6 (in series and in parallel, as appropriate, as in a composite circuit). The inputs of one queue will be the outputs of others; and the possibilities of surges will have important implications on capacity of the regional system. Since, as will be shown in detail in Chapter 6, there is a strongly time-dependent feature in the conditions and demands of the individual air terminal, the same will be at least as true of their composite structure – the regional system. Thus the "steady-state" methods of conventional queueing theory are inapplicable to this problem. This is why we have regarded the questions of capacity, demand, delay, etc., as they apply to the composite regional system, as requiring a new technical attack. During the first year's work on the present contract, we have succeeded in identifying this problem and, by developing methods for the study of the components, have cleared the way for its solution. This would be made during a second year, by methods that are already beginning to take shape.

In Chapter 4, the issues underlying this extension of basic terms and measures and the study of behavior of the subsystems to the full system are examined from a more general point of view, namely, the description of the system in terms of its functions and their related subfunctions.

In closing this chapter, a word on the subject of computer simulation is in order: the fact, namely, that this method, now so popular, has found no place in our discussion of terms and measures. There are two reasons for this omission. The first is the obvious one that the establishment of terms and measures is an act of concept formation: only the mind and not the computing machine forms concepts.

The second reason for the omission of computer simulation – even after the basic concepts have been formulated – can be stated as follows: In order to understand the rational basis of the concepts, to see how the terms and measures work out, general quantitative reasoning must be applied, not merely to one or another special numerical case, but to whole classes of cases. Moreover, the underlying (structural) assumptions in the various cases (in principle, infinitely many) will be different. Computer simulation can give numerical answers in a single case only; or, by varying the input parameters, in cases that all have the same underlying structure (are programmed in the same way). To study as many different structures as are needed for a rational understanding of the terms and measures would require, first, a practically unacceptable number of reprogrammings; and, second, the power of drawing valid generalizations and predictions from sets of numbers. While we see no objection to the use of computer simulation to give an intuitive basis for guessing at theorems that are later verified by mathematical reasoning, we have simply not found such expensive methods necessary.

In contrast to the use of computers for simulation, their use for computation (for which they were originally designed) has formed an important basis for our quantitative results, notably in Chapter 6. In this use, the computer is used to get exact answers by following instructions that are themselves based on mathematical reasoning.

REFERENCES

1. Goldmuntz, L. A., Scanning the Issue, Proc. IEEE, Vol. 58, No. 3, March 1970, pp 289–291; also ATCAC Report (ref. 1 of Vol. 4), p. 3.

2. Aviation Demand and Airport Facility Requirement Forecasts for Large Transportation Hubs through 1980, FAA, Dept. of Transportation, August 1967.

4. DESCRIPTION OF AIR TRAFFIC CONTROL IN TERMS OF SUBSYSTEM FUNCTIONS

4.1 STRUCTURE AND FUNCTION

A regional air transportation system is composed of subsystems, such as the air terminals and intermediate air travel spaces, and the regulatory instrumentalities; these, in turn, are made up of simpler systems, such as runways, control towers, servicing facilities, and many others. To find a valid basis for the definitions of the terms and measures of capacity, demand, delay, and the like, it is necessary but not sufficient to define them for the component subsystems: the system made up from them must be considered as an organic whole of cooperating parts. It is greater than the sum of its parts; and its measure of capacity is not simply found in terms of those of the subsystems.

Actually, our experience with problems of this order has shown us that a model which confines itself to the enumeration of subsystem elements and to describing their physical location and interconnection is doomed to failure, as being basically incomplete: it tends to overlook the functions of the parts and their cooperation in fulfilling those of the whole system. In biological terms, what is needed is the physiology of the system, over and above its anatomy. Therefore, to develop a *systems point of view* – leading to methods whereby the whole can be built up from its parts – we have had to push the analysis of the air traffic control system to the point where it could be described not merely in terms of its physical components, but in terms of its functions, and to see how they are realized by the cooperative interplay of the subfunctions.

The final objective is, of course, the extension to the whole air traffic control system of the concepts of capacity, demand, delay, and so forth, which we have studied extensively in relation to its various component parts (single terminals, and the like).

The analysis of functions into subfunctions, and these, in turn, into still more elementary subfunctions, has been undertaken in a tentative manner during part of the first year of this contract. Although this work has helped us in the clarification of certain ideas, the present state of our results, taken *in toto,* has not reached a fully analyzed form. The parts of our initial steps which we consider sound and in a form suitable for immediate use are described below.

4.2 SIGNIFICANCE OF DESCRIPTION IN TERMS OF FUNCTIONS

As in all scientific applications, a simplified model is needed, showing certain features of concern in the study. In the present project we need a model of air traffic control which will illustrate the idea of "capacity" in the system and major subsystems. The purpose of this model is not to represent the present air traffic control system, or even the improved third-generation system outlined by the Air Traffic Control Advisory Committee. Its purpose is to represent *any* air traffic control system. Such a goal may be unattainable, but we must at least consider a class of air traffic control systems which includes all alternatives under active consideration. On the other hand, we do not need much detail: we simply want to anticipate the kind and amount of service degradation which may result from increasing the amount of various kinds of traffic served.

4.3 THE GOALS OF AIR TRAFFIC CONTROL

The Federal Aviation Act defined the Federal Aviation Administration's mission and objective. The Federal Aviation Administration has, however, considerably wider responsibilities than air traffic control alone, but the mission is defined* as "Ensuring safe and efficient use of the national air space, by military as well as civil aviation, and fostering civil aviation and air commerce." The principal activities of interest that the Act requires in order to satisfy the various statues are: "Air space management and the establishment, operation and maintenance of a civil-military common system of air traffic control and navigation facilities". . . "Development and promulgation of safety regulations including . . . . use of air spaces" . . . "Development of rules and regulations for the control and abatement of aircraft noises" . . . "Fostering a national system of airports; promulgation of standards and specifications for civil airports," and "Formulating long-range plans and policies for the orderly development of air traffic control and navigation facilities."

The system that the Federal Aviation Administration manages, operates, maintains, fosters, and plans shall, in addition, be characterized by safety, economic viability, consistency with national goals (growth and national security), environmental compatibility, user and public acceptability, and self-sufficiency.

*The National Aviation System Policy Summary, DOT/FAA, March 1970.

Thus, the system objective is to achieve safety and efficiency by achieving a compromise between the positive virtues, such as –

- Economy,
- Availability to many users, and
- Convenience,

and negative qualities, such as –

- Degradation caused by multiple air space use,
- Congestion, delay, and collision risk, etc.,
- Degradations, such as noise and conflicting land use, caused by aviation activities to the community at large and to specific groups, and,
- Economic cost of maintaining a system which is capable of allowing the utility of air space to be realized.

4.4 HISTORICAL DEVELOPMENT OF AIR TRAFFIC CONTROL FUNCTIONS

We can translate these general objectives into more specific terms by looking at the historical development of air traffic control. Three origins, in particular,[1] shed light on the goals and purposes of air traffic control:

- Early terminal area control leading to the establishment of the first air traffic control tower;
- Early navigation aids leading to the establishment of airways; and
- Coordination of commercial flight operations leading to the first enroute control center.

In the earliest terminal area control, a man on the ground with visual signaling apparatus augmented the pilot's capabilities by interpreting what he saw and sending visual signals to the pilot. Together with established terminal area flight procedures, this provided a mechanism for avoiding conflict among multiple users of the terminal. Where potential conflict arose, the man on the ground could, within the limits of his vocabulary of signals, direct one aircraft to defer to another, to delay, and to modify his course. He was basically providing a priority rule, time separation, and some space separation. Thus he allowed many users to be served by one strip who might be endangered if each attempted to use the strip with no regard for the time and position of the others. The man on the ground could also assist a landing aircraft in his final approach path and touchdown maneuvers, that is, rudimentary landing guidance, but this was not the essential reason motivating his service.

Terminal area usages and the controller signaling system were improved, standardized, and formalized, and the controller's surveillance was improved by putting him in a tower. The first air traffic control tower was established at the Cleveland Municipal Airport in 1930. Depending, as it does, on visual surveillance by the controller, tower control was available only for visual flight operations for many years.

A second point of origin is navigation. In the early days of aviation, navigation was provided by magnetic compasses and visual reference to features on the ground. Visual light beacons resolved some ambiguity and made certain night operations possible. The invention of the radio range in 1926 made it possible for aircraft to follow a predetermined line without visual reference to the ground. A distribution of light beacons and radio ranges led to the development of networks of airways laid out as straight line segments joining one of these navigational aids to the next. Rules relating flight altitude to direction were promulgated, their effect was to provide altitude separation between aircraft flying in different directions in the same geographical area.

The establishment of these airways had the effect of concentrating traffic directly over the navigational aids, and led to the establishment of special altitude and maneuvering rules for aircraft approaching and receding from intersections. Thus it was already recognized in the 1930's that the ordering of air traffic in itself could produce a concentration not necessarily intrinsic to the concept of air transportation. Aircraft which were not flying on the established airways were given another designated set of altitudes, different from those provided on the airways, so they could fly by with impunity. This was an early example of the joint use of airspace by "cooperative" and "non-cooperative" users.

With this navigational information the pilot reduced his chances of getting lost or of inadvertently flying into mountains, and was able to anticipate his flight path some time ahead. His safety was also improved by the altitude separation rule, which assured that all aircraft flying at his own altitude would be moving in the same general direction, the circumstance most favorable for visual detection and evasive action. In modern terminology, this system provided open-loop, not closed-loop, control, for it had no provision for responsive action based on sensing of an error signal.

The third significant point of origin of air traffic control was the agreement by the commercial airlines flying into the Newark Airport in 1935 to regulate their traffic so that they maintained substantial separation as they traveled the established airways. By this time, each of these aircraft had radio communication with dispatchers on the ground to whom they could relay information about their position. By pooling this information, the dispatchers could determine relative position, anticipate potential close approaches, and redirect the pilots by radio. Initially the system was privately operated by the airlines, and attempted no interference with military and general aviation aircraft, but the system at least provided them with a means of keeping out of each other's way. In July 1936, the Bureau of Air Commerce took over the operation of the three enroute traffic control centers which the airlines had established at Newark, Chicago, and Cleveland. This was a true closed-loop control system, for it provided for the collection of information not available to the pilots unaided and separately, for collating the information and anticipating conflicts (sensing, in control system usage), and for closing the control loop by passing directives and information back to the pilots by radio, causing them to change their behavior in response to the sense data. The use of this scheme is not limited to visual flight operations, for the paths of a well-equipped aircraft can be projected for considerable time and distance by dead reckoning without continuous reference to the ground. Thus, when bad visibility drove nonparticipating aircraft away, the safety of participating aircraft was assured by separation.

A review of modern air traffic control functions as revealed by this historical analysis shows that most of the functions or services provided by air traffic control can be characterized as navigation, separation, or flow regulation. A further service which had its rudiments in all three of the schemes described above is the sensing and relaying of environmental information, from topographical maps to up-to-the-minute wind and weather informatio . It is also possible to use an air traffic control system for other purposes, such as an aircraft early warning system for national defense. With these additions, we can subdivide the three principal functions of navigation, separation, and regulation and develop the following list of principal functions or services provided by air traffic control:

Navigation: coarse (transoceanic, enroute)
Navigation: medium (enroute)
Navigation: fine (terminal area, landing guidance)
Distributing environmental information
Aircraft-to-ground separation
Aircraft-to-hazardous weather separation

Aircraft-to-aircraft separation
Regulation: establishment of route structure and assignment of routes
Regulation: assignment of priority and sequencing
Regulation: gross flow control
Other

Notice that the first six of these functions are necessary in any flight, whether it is served by air traffic control or not. They may, however, be supplied by the aircraft's own instrumentation. Aircraft-to-aircraft separation is necessary only when two or more aircraft fly. Regulation becomes an issue only when many aircraft fly.

4.5 CONSONANCE AMONG FUNCTIONS AND GOALS

It is proper to ask whether all of these functions, which are a legacy of history, still contribute to fulfilling the mission of air traffic control. One extreme view is that only aircraft-to-aircraft separation is an essential air traffic control function. It is true that aircraft-to-aircraft separation ensures safety by preventing midair collisions, and that we have no means other than air traffic control to provide separation when visibility is poor. Recent studies[2] have verified what was believed for a long time, that even in good visibility, see-and-be-seen procedures alone cannot assure separation of high-speed aircraft. Both experience and theoretical studies show that the risk of midair collision is not negligible with present traffic densities. Hence, aircraft-to-aircraft separation is indisputably an essential air traffic control function.

The argument that aircraft-to-aircraft separation is the *only* essential function of air traffic control is hard to sustain. It appears technically possible to supply navigation and other separation functions by, for example, improved aircraft instrumentation, without the control loop implied in the term "air traffic control." Nonetheless, these also serve to help to "ensure safe and efficient use of the national airspace, by military as well as civil aviation, and foster civil aviation and air commerce."

The role of coarse and medium navigation in ensuring safety and efficiency is obvious. Dissemination of environmental information also serves the same ends. It is equally clear that fine navigation in the form of terminal area guidance supports safety, especially if it is effective at night and in poor weather. Separating aircraft from the ground and from weather avoids certain kinds of accidents.

Among the "rules and regulations for the control and abatement of aircraft noise" are route restrictions which can be considered a species of aircraft-to-ground separation where the beneficiary of the separation is not the aircraft but the ground which he is avoiding.

Efficient use of the airspace does not become a problem until constraints are put on the free flight of aircraft by separation requirements, secondary limitations of navigation, the limitation of flight paths for noise abatement, and other consequences of the ways we carry out other functions. However, as things stand now, options are so restricted that a careful juggling of the remaining degrees of freedom is necessary to accommodate the demand. This is accomplished partly by compromises in the ways other functions are carried out. For example, flight paths are laid out not only for minimum length but also for minimum mutual interference.

But regulation, both in the small and in the large, also contributes to the efficient use of airspace: by the purposeful smoothing out of random fluctuations, we can increase average flow rates while decreasing risk and other service degradations. Thus, flow control allows a terminal to operate near its peak rate without building up long queues, and speed-class sequencing reduces certain time losses implicit in random sequencing.

This qualitative review has shown that all of the listed functions, save possibly the undefined "other," can contribute to the achievement of the stated goals and therefore belong properly to air traffic control. Before this project is complete, these qualitative contributions should be turned into quantitative relations, but for the present, a set of qualitative relations which identify interactions is a sufficient basis for further analysis.

4.6 TOOLS FOR FURTHER ANALYSIS

4.6.1 Recursive Pattern of Analysis in Terms of Functions. The identification of these 10 functions and the verification that they contribute to the goals of air traffic control may appear to be an exercise in repetition of the obvious, but it is not sterile. The reader can see that these functions may, in turn, be interpreted as goals for subsystems. What if we then repeat the process, and seek the subsystem functions which support these subsystem goals? (See Figure 4-1.) This will lead us to a structured hierarchy of units of goal and function, in which the function at one level of analysis is the goal

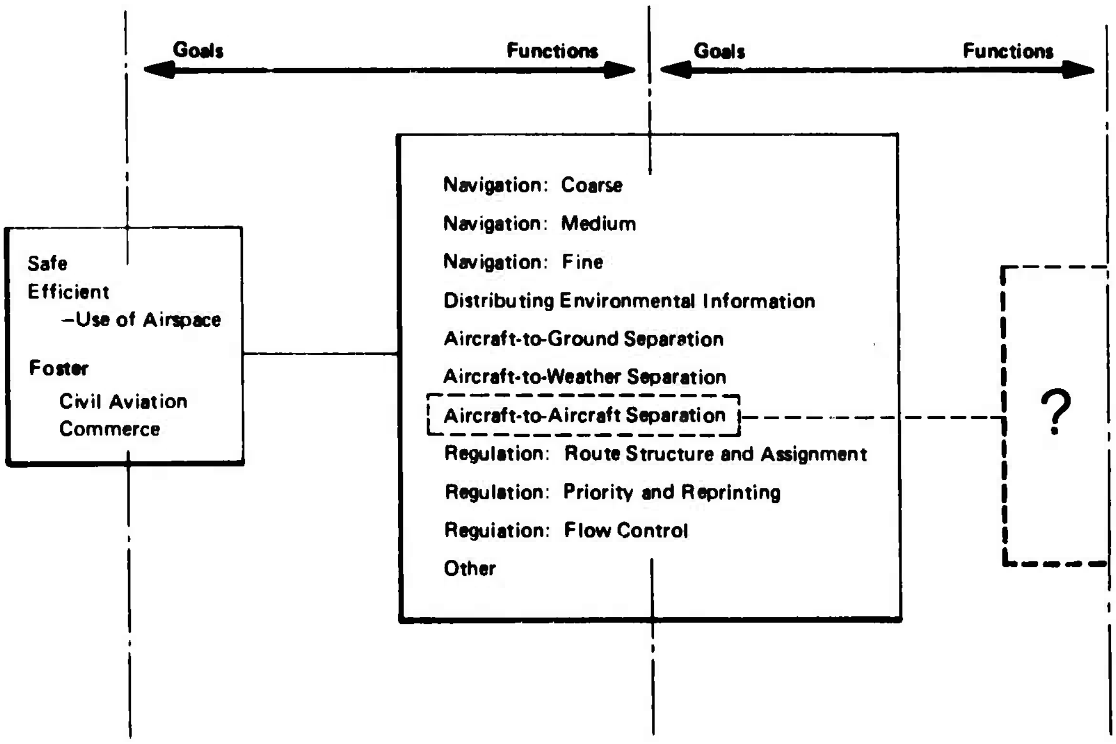

Goals
Functions
Goals
Functions
Safe
Efficient
—Use of Airspace
Foster
Civil Aviation
Commerce
Navigation: Coarse
Navigation: Medium
Navigation: Fine
Distributing Environmental Information
Aircraft-to-Ground Separation
Aircraft-to-Weather Separation
Aircraft-to-Aircraft Separation
Regulation: Route Structure and Assignment
Regulation: Priority and Reprinting
Reguiation: Flow Control
Other
?

FIGURE 4-1 THE HIERARCHY OF GOALS AND FUNCTIONS

for the next lower level. This recursive pattern of analysis is our basic tool for analysis of the system into functions.

Besides providing the pattern for a fundamental analytical method, the steps leading to a list of 11 functions of air traffic control have a secondary benefit of explicitly stating what air traffic control "really is." There is no universally accepted detailed definition. The FAA Gl›ssary[3] describes air traffic control as "a service operated by appropriate authority to promote the safe, orderly, and expeditious flow of air traffic," and the Air Traffic Control Advisory Committee[4] calls it "a service that promotes the safe, orderly, and expeditious flow of air traffic, including airport, approach, and en route air traffic control." We have also alluded to the opinion, on the other hand, that aircraft-to-aircraft separation is the only essential function of air traffic control. When the possible functions have been tabulated, the ambiguity can be resolved by deciding which will and which will not be included.

4.6.2 Distinctions between Information and Material Objects. Air traffic control involves two groups of entities one in the form of information which is gathered, stored, processed, sensed, and transmitted, and the other which involves material objects (aircraft) flowing in real space and time, with a number of links joining elements of these two populations. The content of the information handling part of the whole is mostly symbolic and thus to a degree arbitrary. It is easy to manipulate and to make changes in its design or operation. However, the value of air transportation does not arise from information flow, but from moving the people and things. As far as we can conceive it at present, this will be done in vehicles departing from and arriving at terminals all of which are large, expensive, and relatively hard to change because their principal characteristics are determined by considerations other than air traffic control.

After the distinction has been drawn between material objects and information, it is no longer necessary to insist on a strict functional description of those material elements, the characteristics of which are not defined by the air traffic control system planners. The aircraft, the weather, existing airstrips, the topography – these are all given. Conceptually, we can regard them as modeling themselves, and refer back to their physical reality to abstract its further properties, as analysis requires.

It has been observed that the information handling part of the system carries two fundamental types of data. First, the system obtains information from flight plans on IFR aircraft that expect to fly through the airspace under its jurisdiction. Second, the center receives a flow of radar and beacon data which reflects some of the actualities of the physical flow. It has been asserted[5] that the heart of the air traffic control operation is the reconciliation by the controller of the flight data indicating where the aircraft at any given time should be with the radar and beacon data indicating where the aircraft actually is.

Even in its present rudimentary state, our analysis of functions shows that this description of air traffic control is inadequate. Reconciliation of two data strains in itself does not carry out any of our listed functions, for the element of control is missing. In addition to reconciling the two streams of data, the system must institute purposeful action based on the data stream content. To say that the object of air traffic control is to reconcile these two streams of data is like saying that the orchestra leader's goal is to keep his baton in time with the music, or that the corporate comptroller's goal is to keep the accounts in harmony with the corporation's actual assets and liabilities.

4.6.3 Techniques of Feedback Loop Analysis. Let us take the function aircraft-to-aircraft separation and look at it as a subsidiary goal. The functions which are now used to achieve it are shown in Figure 4-2. From surveillance radar or another source we derive an estimate of the actual position of each aircraft. From the position measurements we sense the distance. If the distance is too small, we generate a control signal, which is communicated to actuators in the aircraft which cause one or both to maneuver. Generating a suitable control signal requires some knowledge of the performance characteristics of the aircraft, and of their relative speeds and aspects. The response of the aircraft to the actuator is rather slow: this and other time delays can be compensated in part by using available data to predict distance rather than to base the control signal on current distance only.

This figure is an example of a feedback control loop. A simple control loop may be schematized in terms of an actuator, capable of more than one action in response to an input signal; a reference signal against which to compare the performance being achieved, a means of comparing the actual performance with a reference signal and sensing the relation between them, and a control signal derived from this sensed relation which controls the actuator. In this case, the actuator is the flight propulsion and control elements of the aircraft, the reference signal is the separation standard, the comparison of actual performance to reference is the distance sensing, and the control signal is, simply, the control signal.

Obviously, many control loops are involved in air traffic control, and we can apply to them the methods and techniques of control system analysis. Without reference to specific equipments, control loops can be described, classified, and characterized in terms of the reference signal (what result is desired?), the comparison sensors (who or what decides what action will be commanded?), and the actuators (what element of the overall complex changes its behavior in response to the control signal?). *A priori,* we know that the delay, frequency response, and noise characteristics of control loops are very important.

Looking back to the aircraft-to-aircraft separation loop of Figure 4-2, we can see some superfluous units: the position measurement is used only to derive a distance estimate. Could aircraft-to-aircraft separation be served by direct (vector) distance measurement, rather than by differencing of two position measurements? This is the principle adopted in recent collision avoidance system (CAS) developments.[6,7]

If the CAS is superimposed on the existing separation loop, the result is two overlapping control loops as shown in Figure 4-3. The conventional aircraft-to-aircraft separation control loop of Figure 4-2 has been simplified, and falls to the lower left, while the CAS loop moves into the upper right. Now, the two sensing functions are using data about the same physical facts, the positions of the aircraft, but they do not receive identical information nor interpret it in the same way. It is quite possible, therefore, for uncoordinated and inconsistent control signals to be generated which lead to instability. The potential instability of nonnested overlapping control system loops is well-known in control system theory.

This simple example illustrates how control loop concepts can be applied to modeling a feature of air traffic control.

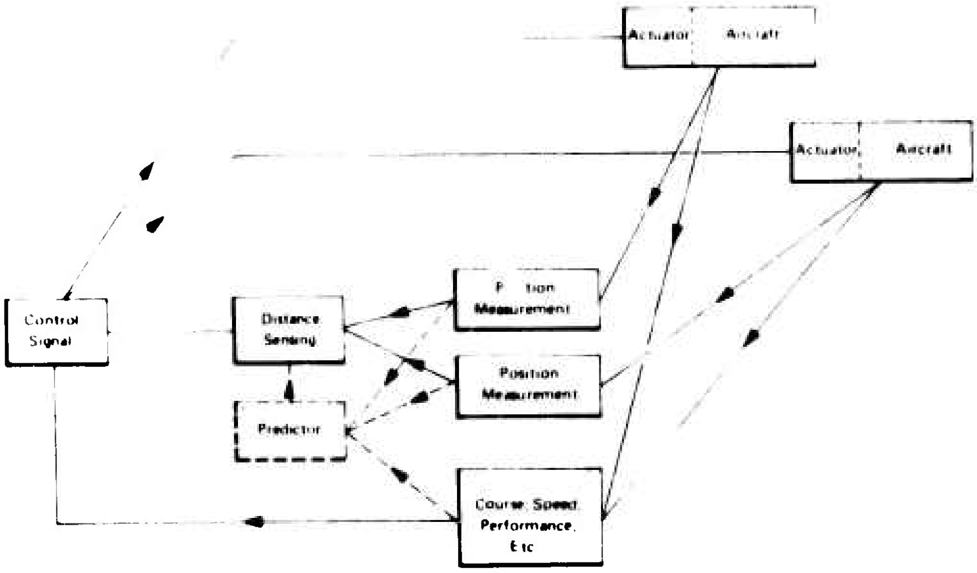


FIGURE 4-2 ONE WAY TO PROVIDE AIRCRAFT-TO-AIRCRAFT SEPARATION

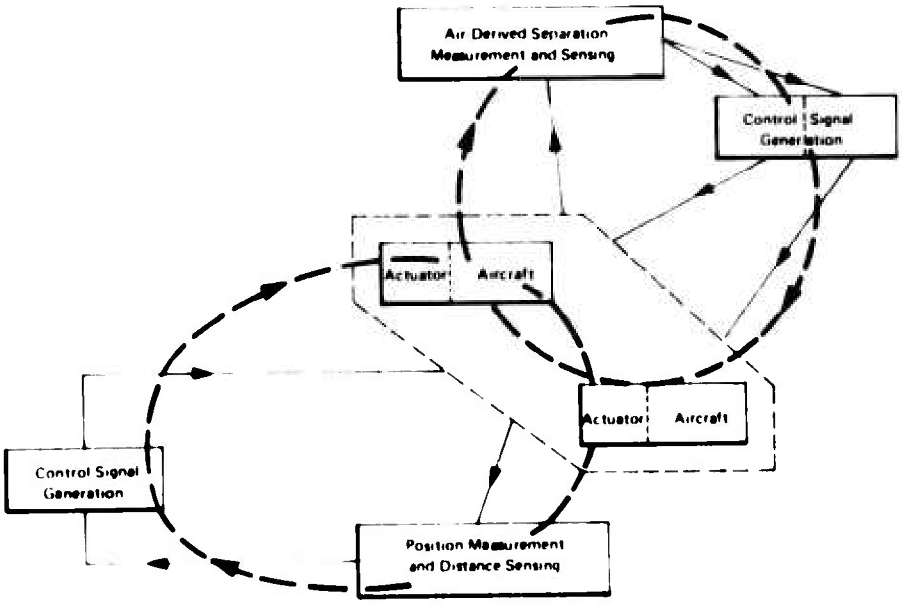

Air Derived Separation Measurement and Sensing
Control Signal Generation
Actuator
Aircraft
Actuator
Aircraft
Control Signal Generation
Position Measurement and Distance Sensing

FIGURE 4-3 TWO OVERLAPPING CONTROL LOOPS RESULTING FROM SUPERPOSITION OF CAS ON CONVENTIONAL AIRCRAFT-TO-AIRCRAFT SEPARATION

4.6.4 Summary. We have shown how we arrived at three tools for modeling the air traffic control system for the purposes of this program:

- The identification of functions required to achieve stated goals, followed by identification of these functions as subsidiary goals, with iteration to produce a hierarchy;

- Separate identification of the properties and flow of material elements of air transportation and the properties and flow of information within the air traffic control system; and

- Application of feedback control loop concepts to air traffic control functions.

The next stage of description of air traffic control system functions has not been carried out to the point which merits presentation in this report, so we shall not go into further detail. We have not yet found any obstacles that would prevent carrying out the next stages, so we remain convinced that an analysis of the air traffic control system in terms of its functions is feasible, and that it will be an important contribution to the methodology of air traffic control system analysis when it is completed.

REFERENCES

1 House report 91-1308, Problems Confronting the Federal Aviation Administration in the Development of an Air Traffic Control System for the 1970's, 29th Report by the Committee on Government Operations, July 16, 1970; also private communications from Russell Biermann, and others, FAA.

2 Graham, W., and Orr, R. H. Separation of Air Traffic by Visual Means: An Estimate of the Effectiveness of the See-and-Avoid Doctrine, Proc. IEEE, Vol. 58, No. 3, March 1970, pp 337 360.

3 FAA Glossary, FAA Handbook 1000.15, June 15, 1966, revised August 18, 1967.

4. Report of Department of Transportation, Air Traffic Control Advisory Committee, Volumes I and II, U. S. Government Printing Office, December, 1969.

5. House Report 91-1308, op. cit., p. 67.

6. Report of Dept. of Transportation, op. cit., Vol. I, section 3.3.3, Air Derived Collision Avoidance System.

7. Borrok, M. J., and Rider, D., Results of the ATA CAS Flight Test Program, presented at Session IV, CAS, National Air Meeting on Air Traffic Control in the 1970's, The Institute of Navigation, St. Louis, April 16, 1970, also other papers of Session IV.

5. SAFETY

5.1 THE RELEVANCE OF SAFETY TO THE ANALYSIS OF AIR TRAFFIC CONTROL SYSTEM CAPACITY

In general terms, the interrelationship between safety and capacity in air traffic is widely recognized. For example, the Air Traffic Control Advisory Committee report[1] says, "The Committee concentrated on control of aircraft through the air space, from takeoff to landing. Emphasis was placed on the denser portions of the air space where the danger of mid-air collison and the need for efficient use of scarce resources (principally runways and terminal air space) make sophisticated ATC mandatory if safety is to be assured without sacrifice of capacity and without unacceptable delays or interference with freedom of flight." Marner[2] makes a similar point, including the statement, "For a given air traffic control system, safety and capacity are implicitly related." It is not difficult to find many such statements. Nevertheless, because we are putting so much emphasis on safety, we want to examine the basis for this conclusion.

Air traffic control system capacity is really not a single number; it is a complex relationship among a number of variables. Some of the philosophical aspects of this conception have been discussed in Chapter 3. Here we shall show their bearing on definitions of safety.

Let us imagine for a moment that we can draw up a complete description of the air transport system as it functions in a particular set of circumstances. Suppose that we have specified the environmental information and other variables over which we have no control, the various operating parameters, and other variables, probability distributions, and functionals necessary to specify how the system is working. Imagine also that we can determine whether the overall operation of the system is satisfactory or unsatisfactory. If we regard the variables which describe the system performance as coordinates in a space (probably a space of infinite dimension), then we have divided this space into regions of satisfactory and unsatisfactory operation.

Risk, of course, actually figures in the specification of these boundaries. Consider the following conceptual experiment. Take any satisfactory operating condition of an air transportation system, and keep all parameters fixed except for risk. Imagine that the risk is gradually increased without any effective change in the other operating parameters. Obviously, if the risk is increased enough, the overall performance will be considered unsatisfactory. We can therefore conclude that a change in risk is sufficient to distinguish in certain cases between satisfactory and unsatisfactory operation. Therefore, some parameter related to risk must be involved in the specification of the boundaries.

A refinement of the same argument shows that a good balance of all values and costs will never result in negligible risk. For if the risk is negligible, we could make the risk a little bit greater, say, by increasing the flow rate a trifle and decreasing separation standards, with a net increase in "capacity." We are only restrained from this course when risk exerts a finite influence, which means it is no longer "negligible."

It might still happen that the various rates of flow which characterize the throughput of an air transportation system are relatively insensitive to risk. This, however, is most certainly untrue. For example, Goldman,[3] Astholtz et al,[4] and Steinberg[5] show how sensitive the runway arrival capacity is to separation standards, and how tightly coupled separation standards may be to risks in the landing operation.

This strong dependence of terminal area flow rate on safety contradicts the general observation that safety has not decreased with an increase in air traffic. For example, in the 13 years from 1953 to 1966, the number of passenger miles flown in domestic scheduled air transport planes more than tripled, but the passenger fatalities per passenger mile appeared, if anything, to decrease slightly.[6] But this contradiction is not real.

In the first place, over a period of time other influences motivate us to reduce risks at the same time that we increase the amount of air travel. We do not have an equilibrium condition in which costs and benefits achieve, once and for all, a static balance. The annual cost of aircraft accidents is in the neighborhood of $1 billion,[7] a figure large enough to represent a considerable constraint.

Furthermore, risk is treated by many people as an inelastic constraint. For example, Steinberg[5]* makes an allusion to "nominally acceptable" safety. Holt and Marner[8]** refer to

* Page 315.

** Page 370

"the tolerable level of passenger, crew, and 'controller' concern." Hagerott and Weiss[9] refer in several places to "an intolerable hazard. . .acceptable probabilities of not violating separation minima. . .the tolerable level of concern of the passengers, crew and controllers," and use other similar expressions. This suggests that the risk is treated as an intermediate parameter implicitly stabilized at some acceptable value and serving as a fixed fulcrum for the balancing of other parameters. Ratcliffe[10] has stated that "when faced with a serious overload sit ation, the controller normally preserves air safety and his own sanity by slowing down the traffic demands, by one means or other."

From our own observations, we infer that the risk of fatal accidents and other serious service degradations influences the system behavior in at least two distinct ways. Controllers and pilots make and carry out decisions which, within minutes, hold down the rate of flow and transform or convert immediate and severe performance penalties into less severe ones. In other words, their collective decisions and actions reduce the risk of fatal accidents, accidents with injury or physical damage, and other acute penalties at the risk of increasing delays or fatigue to personnel and other less acute penalties. Also, the cumulative effect of service degradation over a long period is to reduce demand. Carriers and general aviation will avoid operations in congested areas if their needs can be met in part by operations elsewhere, and passengers and shippers learn to avoid flights where the risk of delay or cancellation detracts too much from the value of air transportation.

Note that successfully holding down the fatality rate and keeping the rate of cancellations and reroutings to a level where their economic impact is acceptable results in more delays and creates the illusion that delay is the principal penalty for exceeding capacity. But adding another aircraft to the system cannot in itself cause delay. At worst, adding an aircraft decreases safety and increases workload. The delay results from the operation of the air traffic control system itself, and results from an implicit trade among other performance variables intended to maintain safety. The choice of delay as the currency with which to pay the debt incurred by threatened overloads is a choice, the nature of which cannot be understood without knowing explicitly the trades available among all forms of disruption.

The importance of such trades can be emphasized with a simple example. Suppose technical improvements in radar reduced the time lag to display evidence of an in-air conflict to the controller, and technical improvements in communications cut down the time required to transmit a correcting directive to one or both aircraft. Intuitively, we can see that these improvements should increase the "capacity." By how much? If the momentary bottleneck is controller workload, we may be able to estimate directly the controller's ability to make more judgments in exchange for quicker communications. But if the bottleneck is airspace, the increase in capacity will come from reducing separation standards.

We need to know how much the separation can be reduced so that, with these technical improvements, the users are as safe as they are now with current technology and procedures. Thus, we believe, in general, no definition of, formula for, or method of estimating capacity will have any predictive value unless it treats safety also.

Furthermore, the effects that these mechanisms have on constraining and shaping future demands are often overlooked. If air travel were absolutely free of risk, if its cost were negligible, if it were instantaneous – in short, if every man had a magic carpet on which he could wish himself anywhere – the demand for such travel would grow tremendously. The demand for air travel does not limit itself – something limits it. The long-term demand can be stabilized by increasing the cost of service or by degrading the value of service rendered. The restraining influences may be explicit or hidden, they may spread the cost equitably or capriciously, and their actions may be random or predictable. Therefore, although we may treat degree of risk as a rigid constraint for analyses and suboptimizations on a small scale, it is a fundamental error to do so in the long run.

Quantitative relations between risk and air traffic control system parameters are not usually stated explicitly. It is extremely difficult to docun.ent a negative statement of this kind. However, some evidence can be adduced. For example, a compilation of reference material for air traffic control separation[11] is intended to be a "common reference document regarding existing aircraft separation to be used in interviews, task discussions, and for updating (currency of) operational input to the ATC Advisory Committee." In 141 pages of text and figures, we found no quantitative reference to risk. Steinberg[5] suggests some quantitative relationships, but states in his conclusions that more quantitative information is required. Marner[2] makes such statements as "quantitative safety goals could be established and various system parameters could be traded off with various capacities. This may seem overambitious, but I believe we now have the techniques for accomplishing this," and other statements from which we can infer that he believes that satisfactory quantitative relationships have not yet been established. The most commonly cited sources with quantitative estimates appear to be several reports by Reich[12, 13] and a report by Marks.*[14]

One reason for this lack is probably simple ignorance. It is difficult to produce convincing logical statements in which the amount of risk figures, and it is even more difficult to do experimental work to substantiate them. We shall see** in Appendix A that direct observation of fatal accidents will not produce enough evidence upon which to base useful statistical inferences about the relative safety of methods and procedures. This is quite unlike the situation with automobile traffic safety, where the number of events is large enough to yield significant statistical results over a short period of time. Therefore, not only are we faced with a difficult problem, but one in which the measurements must be made indirectly.

Furthermore, there is considerable reluctance to talk about hazards in the air transportation community. Like other negative observations, this one is quite difficult to substantiate. We have come away from many meetings with the feeling that reference to risk in air transportation is considered to be in bad taste, just as references to death by cancer and tuberculosis were a generation ago in drawing room conversation. The aforementioned reference material[11] avoids reference to risk altogether. The ATCAC interview guide questionnaire[16] has several questions on risk at the very end of the questionnaire, but the interviewer's instructions explicitly command him to take the questions in sequence and not to return to an earlier question after leaving it. Can we infer that these questions were an afterthought, or that the subject of safety cannot be raised until all other subjects have been covered?

It is even possible that some of the things which are widely believed about the relationship between safety and air traffic control system parameters may not be completely true. For instance, we have been told a number of times that the 3-mile separation standard imposed between two aircraft following one another for a landing is imposed widely to avoid mid-air collisions, and that pilots are very reluctant to accept a smaller separation in IFR weather.*** The situation in which several aircraft proceed in a chain at near minimum spacing is very common, occurring sometimes for hours at a time, day after day, at the busiest airports. Yet, in the last decade, there have been no collisions in the United

*Cited by Refs. 5 and 12.

**Also see Reference 15.

***In some circumstances, jet vortices also limit minimum spacing, but that is another matter.

States between two carrier aircraft under simultaneous radar surveillance and simultaneous control by one controller.*

This may be evidence of good management or just good luck, but it may also be that 2-mile or 1½-mile separations are simply not so dangerous as pilots and controllers believe them to be. The risk as a function of minimum separation standards may rise sharply over a narrow range of possible standards, or it may be a very gradual function – that is, the constraint imposed upon the separation standard by risk may be relatively inelastic or elastic. Ratcliffe[10] suggests that the risk may be subjectively evaluated by the controller in terms of his own workload, and that he imposes larger separations to thin out traffic and thus to relieve his anxiety. Experience in formation and cluster flying, and specialized operations, such as aerial refueling, suggests that with good sensing and appropriately organized control loops, spacing very much closer than 3 miles is compatible with at least moderate safety.

Even if experience were reliable, it fails as a guide in extreme or totally new circumstances. For example, in the regimes suggested by Astholz et al[4] or McFadden,[17] circumstances are so changed that presently accepted separation standards may be quite irrelevant. Unless we test the safety analytically or empirically, we may deprive ourselves of capacity benefits arising from such new developments as precision navigation equipment and proximity warning equipment.

Once again, we conclude that a quantitative relationship between risk and operating parameters is necessary for an understanding of air traffic control system capacity.

Our conclusions relating to the relevance of safety to air traffic control system capacity may be summarized as follows:

- Reduction or bounding of risk to human life is an important goal in many ATC planning and operating decisions;
- Quantitative relations between risk and the ATC system parameters are seldom stated explicitly, partly because of reluctance to deal openly with the issue of fatality and partly because so little is known about them;

*In the collison over the Grand Canyon in 1956, there was no radar coverage, and in the collision over New York in 1960, the two aircraft were responding to different controllers, although one of them drifted away from its holding pattern.

- More understanding of the relation of operating parameters to risk is necessary before a quantitative measure of capacity can be defined and usefully applied to the ATC system and its major subsystems.

5.2 DEFINITIONS OF SAFETY IN AIR TRANSPORTATION

The impact of aircraft accidents can be analyzed in two stages:

1. How many accidents occur and with what number of fatalities and injuries and what kind of damage; and

2. What is the cost of each fatality and each kind of damage and injury.

In rational analyses involving compromises among many costs and benefits, we may regard safety either as a constraint or as a variable to be traded. With safety as a rigid constraint, it is not necessary to reduce the various measures of the cost of accidents to common units. However, if risk is one of the variables to be traded, we must either reduce all values to a common unit, such as the dollar, or deal with the complexities of a value system with two or more incommensurable units. For analytical simplicity, it is highly desirable to reduce all costs and benefits to a single unit, and this is the course traditionally adopted by economists such as Fromm.[7] Many people, including, for example, Schelling,[18] believe that expressing the value of human life in dollars overlooks some components of value. Fortunately, as we shall show, this dilemma can be partially alleviated in at least two ways. First, many comparisons and subsystem analyses can be completed with an analysis of numbers and kinds of accidents without necessarily stipulating the dollar cost; second, it is possible to make definitions of safety which can be used in trade-off studies and which compare risks not to dollars but to other risks.

No matter what measure we use, loss of human life is the most costly consequence of aircraft accidents. Using only most direct dollar costs, Fromm[7] estimates* that accidents account for three-quarters of all of aviation support ineffectiveness costs, and that fatalities account for more than five-sixths of the cost of accidents. With human fatalities accounting for over five-eighths of the total aviation support ineffectiveness costs, in economic terms alone, the cost of human fatalities dominates all other terms, and in a preliminary discussion of safety we can limit ourselves exclusively to fatalities.

*See Tables VII-2, -5, and -6 of reference 7.

The question of how to place a dollar value on a human life has largely been answered, and in the context of air transportation this answer is laid out in detail in reference 7. This is not to say that the answer is complete or fully accepted: Schelling[18] has expressed some divergent views, and the discussion provoked by his exposition illustrates the degree of controversy which still persists. Much of the uncertainty concerns the accounting of the value of human sensibilities and emotion and the costs of uncertainty which would add additional terms to the straightforward economic elements outlined by Fromm. Altogether, over and above the reckoning of reference 7, these would increase the cost of an airline fatality and attribute to fatalities an even larger proportion of the total accident costs.

Fortunately, for many purposes economic costs need not be determined. Other things being equal, we know that the safety measure which reduces the probability of a fatality 20 percent is better than one which reduces it only 10 percent.

Nevertheless, as soon as predictions or observations of a number of accidents are to be studied, a question of normalization or of units arises. This can be illustrated with an analogy from mechanics. We all use the words – work, power, force, and pressure – and know how to distinguish their technical meanings. In the technical sense, each is quite distinct from the others, and our use of each term is accurately supported by intuition and past experience. Nevertheless, they can be viewed as four normalizations of the same measurable quantity: using work as the fundamental unit, then power is work per unit time, force is work per unit distance, and pressure is work per unit volume.

For the same reasons that we use a variety of normalized units related to work, in air transportation we may wish to use a variety of normalized terms relating to fatality. The most common unit, and the one normally used in comparing different modes of transportation is the fatality rate per passenger mile.[6,19] For other purposes the rate of fatal accidents per departure has been tabulated.[7]* The ATCAC report[1] notes that no single measure of accident or fatality rates is satisfactory for all purposes, and adds[1]** a third normalization to get fatalities per passenger hour or per aircraft hour flown.

*See Table V-9 of reference 7.

**Volume 1, p. 17, of reference 1.

There are many special relations among these measures. For example, the number of fatal accidents per hour of operation equals the number of fatal accidents per mile of operation times the average speed. The number of fatal accidents per departure equals the number of fatal accidents per hour times the average duration of a flight leg. The number of fatal accidents per departure equals the number of fatal accidents per mile times the average length of the flight leg, and so forth.

There are some other considerations special to certain categories of accidents. For example, as a general rule an aircraft involved in a mid-air collision usually either crashes and kills all occupants or lands with no fatalities. Hence, for this particular category of accidents, the risk in a given operation is the same when expressed as the probability of a fatal accident per mile of vehicle operation or the probability of passenger fatality per passenger mile. However, integrated over a heterogeneous population, the overall risks may be different.

To take a numerical example, suppose we have 1 billion miles of operation of general aviation aircraft carrying five people each with a mid-air collision rate of 1 per 100 million miles, and 1 billion miles of operation of air carrier aircraft carrying 50 people, each with a mid-air fatal collision rate of 1 per 1 billion miles. Thus:

Class of Aircraft	Plane Miles	Passenger Miles	Total Accident Rate	No. of Fatal Accidents	No. of Fatalities	Fatal Accidents / Plane Mile	Fatalities / Passenger Mile
General	10^9	$5 \cdot 10^9$	10^{-8}	10	50	10^{-8}	10^{-8}
Air Carrier	10^9	$5 \cdot 10^{10}$	10^{-9}	1	50	10^{-9}	10^{-9}
Total	$2 \cdot 10^9$	$5.5 \cdot 10^{10}$	NA	11	100	$5.5 \cdot 10^{-9}$	$1.8 \cdot 10^{-9}$

The total number of plane miles is 2 billion, the total number of passenger miles is 55 billion, and the overall fatal collision rate per plane mile is 0.55 per 100 million vehicle miles, while the fatality rate is only 0.18 per 100 million passenger miles.

Another hypothetical example, leading to a paradox, has been expressed by Fromm:[7]*

> *"Approximately 70 percent of air carrier accidents take place in the terminal area and are incident to take-off and landings. Thus, if the number of departures and accident prevention efforts are held constant while the average length of the trip is increased substantially, the accident rate will appear to be falling dramatically even though no corrective safety actions have been taken."*

* Pages V-22, V-23.

From these examples, we can see that definitions of risk which appear to be equivalent under certain very plausible conditions may become quite different if parameters such as the mix of aircraft or the average length of a flight are varied. Therefore, in the present study we have to recognize such distinctions, including some which may have appeared as sophistries in the past, and point out those which may become important in later stages of a comprehensive analysis of air traffic control system capacity.

In a previous discussion of the relevance of safety considerations to a study of capacity, we pointed out that risk is sometimes regarded as a rigid constraint; that is, one assumes the existence of a degree of risk which is "acceptable" or "tolerable." We have just seen how the relative risks in certain situations may be made to appear greater or smaller according to the normalization of the risk unit. This invites the question: If there is a tolerable threshold of risk, in what units is it to be measured?

Ultimately, death is inescapable. Every human activity has some risk of death associated with it: one may slip in a bathtub, suffer concussion or unconsciousness, and drown; one can choke on a piece of food while eating; one may be run down by a passing automobile while crossing the street. In our society, most people are made aware of the order of magnitude of many risks in everyday life. We are told how many years of life expectancy has been added by modern medical science to an infant born in our society. The Surgeon General's report tells us the burden to our life expectancy which we impose by heavy smoking. Our life insurance agents translate probabilities of survival into premium charges, which most of us are ready to pay. Anyone who takes the trouble can find out that his probability of dying in the next year is greater than 1 in 1000, more like 1 in 200 for mature adults in good health, and considerably greater for the ill, the elderly, and those who live hazardously.

In the face of this reality, we could expect that below some level of incidence the perception of an additional risk of death would be masked by our uncertainty of survival expectancy due to existing causes. In terms pertinent to air travel, suppose that in some sense the risk of flying were small compared to the risk of contracting a disease and dying. Would this be perceived as a "small" risk?

When the question is phrased this way, it is clear that the appropriate unit of risk is expected number of fatalities per hour of exposure of the subject. This measure has been studied by Starr,[20] with positive results. With risk measured by the expected number of fatalities per hour of exposure of the subject, he makes the following points:

- The amount of risk, as so defined, which people are willing to accept correlates well with the amount of benefit or pay they expect to get while exposing themselves to it;
- The proportion of the population willing to participate in any activity correlates inversely with the risk;
- These relations hold over a wide range of human activities, including hunting, smoking, automotive travel, and fighting in Vietnam, as well as both general and commercial aviation; and
- The threshold where most people are willing to participate for a small perceived benefit is near the level of risk of contracting a fatal disease, which is around 1 fatality in 1 million hours of exposure.

Figure 5-1 (reproduced from reference 20) illustrates most of these relations. The vertical axis is risk P_f in fatalities per person-hour of exposure. The horizontal axis is average annual benefit per person involved, converted to dollars. Starr admits that this conversion is the most uncertain aspect of the correlations. The names of a number of activities or sources of risk, such as general aviation, Vietnam, and commercial aviation, are placed on the figure at points, the coordinates of which are their respective benefit and risk.

The figure also shows two stippled areas labeled, voluntary and involuntary, which represent the transition region from unacceptably high risk at the upper left to acceptable risk at the lower right. Voluntary exposure is one which the subject may choose to avoid, like smoking. Involuntary exposure is one which the subject may not easily choose to avoid, such as hurricanes and other natural disasters. The transition from unacceptable to acceptable risk is not sharp; as this zone is crossed, Starr finds the proportion of the population that will participate and therefore accept the risk runs from near zero to a large fraction.

For reference, the probability of contracting a fatal disease is also plotted, about 1 per 1 million hours of exposure for the whole population, and about 1 per 10 million hours of exposure for the military age group.

If we observe that most of the time a fatal accident in an aircraft kills everyone aboard, we can translate this index into a measure of vehicle safety:

Risk = probability of a fatal accident per hour of operation.

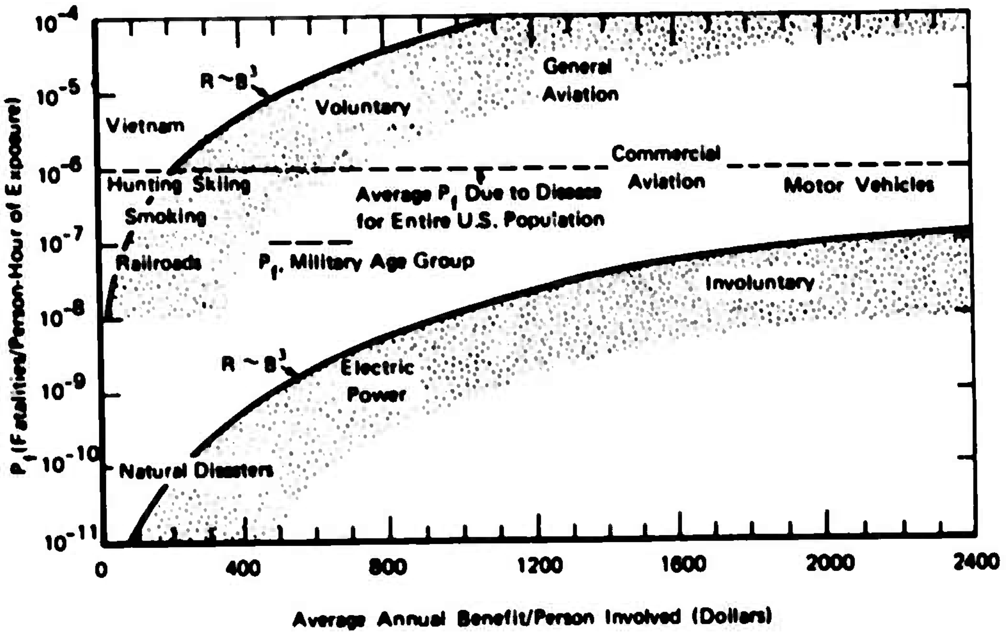


FIGURE 5-1 RISK (R) PLOTTED RELATIVE TO BENEFIT (B) FOR VARIOUS KINDS OF VOLUNTARY AND INVOLUNTARY EXPOSURE (From Starr, Ref. 20)

The current risk for general aviation is around 20 fatal accidents per 1 million hours of operation. This is considered dangerous by many people. For air carriers it is around 2 per 1 million a bit above the threshold and considered somewhat risky by quite a few people. For riders in automobiles, it is about 1 fatality per 1 million hours of exposure. This appears to be about the threshold required for nearly 100 percent participation. The risk is much lower for passengers in buses and trains, which are considered very safe by nearly everyone.

This figure and the rest of Starr's thesis should not be interpreted to show causes; this is merely an example of descriptive statistics. We cannot yet say whether high risk discourages participation, or whether high participation demands lower risk, nor can we say whether a high perceived benefit leads to investment resulting in safety, or whether high risk makes the subject demand a high return.

If we accept the principle that an individual has a quantitative appreciation of the risk of death from all causes, that this total risk is in some sense tolerable, and that an additional risk which increases the total risk by a small proportion will be perceived as small, then we may have found a basis for defining in quantitative terms a tolerable level of risk in air transportation. Even if the numerical value of the threshold is uncertain, we have made progress toward defining the unit of measurement.

The terms of this measure of risk are consistent with the simplified model of a user deciding how to allocate his hours with minimum risk and maximum benefit. Such a model may be unrealistic. Often, duration is not a critical mission parameter. In travel, the choice is more often how to travel for x miles than it is how to travel for x hours. Yet, the availability of air transportation enables a person to travel much further than he could have travelled on trains and boats. It is clear that people take advantage of this opportunity: instead of merely traveling the same distance that we used to generations ago in less time, we travel longer distances.

Our conclusions may be summarized as follows:

- Inasmuch as loss of life contributes most to the total cost of aircraft accidents, aircraft safety can be discussed initially in terms of fatal accidents.

- There are at least six useful measures of risk, in three pairs:

  1. Fatalities per person mile, and

  2. Fatal accidents per vehicle mile.

(These are particularly useful in comparing risks of competing means of transportation serving the same purpose.)

  3. Fatalities per person departure, and

  4. Fatal accidents per aircraft departure.

(The usefulness of these measures is special to air transportation and arises from the observation that most fatal accidents occur during landing, take-off, or in the immediate terminal area.)

  5. Fatalities per person exposure hour, and

  6. Fatal accidents per vehicle operation hour.

(These appear to have special usefulness in defining safety objectives with reference to socially acceptable risks and independent of comparison with other transportation modes.)

5.3 KINETIC MODELS OF AIR TRAFFIC SAFETY

We have done some exploratory analysis using kinetic models similar to those in references 21 and 22 to gain some familiarity with the statistics of mid-air collisions and near encounters. These models assume that a number of particles are moving independently and randomly in a region, with some known distribution of direction and velocity. Because of the assumption of randomness, we are able to compute the probability of collision or of near-encounter within a stated distance.

It should be noted that such an analysis is mathematically equivalent to a problem in search theory, the only difference being that in search theory an encounter is regarded as a favorable event whereas in near mid-air collision theory an encounter is regarded as an unfavorable event. Setting aside this value judgment, we can invoke all of the mathematical resources of search theory, the systematic theory of which dates back to World War II.[23]

A frequent question in search theory is where to look for an aircraft (or a ship, or a submarine, or a guided missile) which may issue from one of a number of starting points toward one of a number of targets, with considerable discretion in the choice of paths in the region in between. An extremely common result of such analysis is that the most favorable areas in which to search are the areas immediately surrounding the starting points and the terminal points, whereas search in the area intervening is relatively unfavorable. The reason for this result is that the probability density distribution may be uniform and rather low in the intervening spaces, but must peak at both the starting points and the end points.

This suggests that the probability of mid-air collisions or near mid-air collisions among aircraft probably peaks near terminals and is rather low far away from terminals. This result is well known, and has been reduced to fairly quantitative form by Steinberg.[5] We can ask the additional question: How much of the risk of mid-air collision is due to the nonuniformity of distribution of air traffic? Or, to put it another way, if air traffic were uniformly and randomly distributed over the United States, and if everyone flew randomly and blindly, what would be the probability of mid-air collision?

Because this is to be a rather rough-and-ready calculation, let us use the simplest model, that of reference 22. We start with the formula

$$L_2 = \frac{n}{2r\sigma} \qquad (5\text{-}1)$$

where L_2 is the mean distance an aircraft travels without a collision, n is the effective number of altitude layers, r is the average minimum approach distance between two aircraft which will result in collision, and σ equals the total aircraft density per unit ground area. From reference 24 we take the figure for average collision cross section:

$$A = 2rh = 1670 \text{ sq ft} \qquad (5\text{-}2)$$

where A is the average collision cross section, construed as a rectangular window of height h and width $2r$. If we assume that the maximum effective flying altitude is H, then we can write

$$H = nh \qquad (5\text{-}3)$$

from which it is easy to derive

$$\frac{n}{r} = \frac{2H}{A} \qquad (5\text{-}4)$$

and

$$L_2 = \frac{H}{A\sigma} \tag{5-5}$$

Now let us do a numerical example. From reference 6 we discover, for the year 1967, that certified air carriers in domestic service flew 1,462,240,000 miles in a total of 4,136,347 hours of revenue operation. In the same time, general aviation flew about 3,440,000,000 miles in 22,150,000 hours, giving a total of 4,902,000,000 miles of operation in 26,286,000 hours. If these operations are uniformly distributed through 8760 hours of the year, the mean number of aircraft in the air at any one time is almost exactly 3000. The average speed in the air is 187 miles per hour.

The area of the United States is about 3,615,000 sq miles. From this it follows that the density is:

$$\sigma = \frac{3000}{3{,}615{,}000 \times (5280)^2} = 2.99 \cdot 10^{-11} \tag{5-6}$$

and, assuming that the maximum operating altitude H is approximately 30,000 feet, it follows that

$$\begin{aligned} L_2 &= \frac{30{,}000}{1{,}670 \times 2.99 \cdot 10^{-11}} \\ &= 6.0 \cdot 10^{11} \text{ feet} \\ &= 114{,}000{,}000 \text{ miles} \end{aligned} \tag{5-7}$$

Using the fact that the average air speed is 187 miles per hour, we infer that

$$\begin{aligned} \text{mean time between collisions} &= \frac{114{,}000{,}000}{187} \\ &= 610{,}000 \text{ hours} \end{aligned} \tag{5-8}$$

and, hence, that the expected number of collisions in a year is

$$\text{expected number of collisions} = \frac{3000 \times 8760}{610{,}000} = 43 \tag{5-9}$$

The actual number in the year 1967 was 27 (reference 25).

Thus we see that if all the aircraft in this country had been flying around in paths randomly distributed in space and time, they could have flown blindly with only a slightly greater incidence of mid-air collisions than was actually experienced. It is patently clear that we exert a great deal of effort to keep the actual accident rate down as low as this. Therefore, either our efforts are ineffectual or the assumption of random distribution is insufficient. Of course, it is the randomness assumption which is at fault. In fact, the distribution of air traffic in both space and time is highly nonuniform. But this allows us to confirm the intuitively obvious fact that mid-air collision risk is a problem of peak hours and high density routes. Furthermore, because of the structure of our ground complex in support of air transportation, there is no hope of reducing the high concentration of traffic near the air terminals of metropolitan hubs, and so the problem of collision avoidance in terminal areas will remain.

We could be a little more precise by using the method of reference 21, which uses the velocities of both members of a pair to compute a mean *relative* velocity $\overline{V}_r$. The value of the integral expression (reference 21, Figure 2) for $\overline{V}_r$ has been published in reference 23, the author of which attributes the formulation and evaluation to the late George Kimball. It is:

$$\overline{V}_r = \frac{2}{\pi} (V_0 + V_1)\, E(\theta) \qquad (5\text{-}10)$$

where E is a standard elliptical integral of the second kind, and

$$\sin \theta = \frac{2\sqrt{V_0 V_1}}{V_0 + V_1}$$

Note that if we define

$$V_{max} = \max(V_0, V_1) \ , \qquad (5\text{-}11)$$

then

$$V_{max} \leqq \overline{V}_r \leqq \frac{4}{\pi} V_{max} \qquad (5\text{-}12)$$

in any case, so an effective estimate of fair precision is always achieved by replacing $\overline{V}_r$ by the larger of the two velocities. From this it follows that an estimate based on the average velocity underestimates the accident rate, but not by a great deal.

5.3.1 Probabilities of Intrapath Collisions in Blind Approach. In another instance, we have examined a one-dimensional model encounter which is representative of collisions between succeeding aircraft following the same path to a landing. When aircraft are approaching an airport in conditions of heavy traffic, and when the visibility is so low that they depend on air traffic control radar to make their approach to the runway, they are confined to a restricted path in space and a definite minimum spacing along this path. The approach path is usually curved, spiraling down to the neighborhood of the runway from which the visual landing is made. The length of the path is very much greater than its lateral dimensions, and it has been compared with a winding piece of spaghetti in space. If all electronic equipment is functioning well and the pilots obey orders, the only reason for two aircraft on the same path to collide – to produce, as we shall say, an *intra-path* collision – is the irreducible inaccuracy in radar positions, the added imprecision in conveying position information to pilots (partly due to an irreducible random delay), and, finally, in the inaccuracy with which the pilot can direct his aircraft to comply with instructions.

The method for countering the danger of collision due to these (and other) unavoidable inaccuracies in carrying out the ordered plan is to avoid too tight coils of the spaghetti path, and also to maintain a spacing along the path which does not fall below a certain minimum value S. This, however, is expensive in delays and airport saturation. The object of this subsection is to make a quantitative evaluation of the increased probability of intra-path collision incurred by lowering S – and thus increasing airport use. The problem will be formulated in terms of a simplified model, intended to represent the main quantitative interreactions.

The first step in this simplification is to treat the position of each aircraft on the track as given by one variable; the spaghetti is replaced by its central curve C, and the position of the i'th aircraft is given by the arc-length s_i measured along C *from* the landing point *to* the aircraft. The event of present interest is that of two different aircraft (i and j) moving so that their arc-lengths coincide (s_i-s_j). While it is recognized that this could occur without the aircraft physically colliding – as they might have sufficient lateral separation along their spaghetti, or they might see each other in time to dodge – certainly the event of $s_i = s_j$ is one of great danger, the probability of which is very important to know and control.

The second step in the simplification deals with the time evolution of the system of aircraft moving along C toward the landing point. We separate the motion of each into a uniform part common to all of them and having a constant mean velocity made good -v (speed $v > 0$ toward the landing point), and into random individual changes in position. Then a moving reference system is introduced on C (thought of as straightened out), moving with velocity -v. This has the effect of *subtracting* -v (since $v > 0$, of *adding* a positive quantity) to the (negative) velocities of each aircraft. Their positions can then be represented by points executing their individual random motions about their assigned points (at the S spacing), now all *fixed*. Another effect of relative motion is that the landing point is replaced by a reference point moving up C with speed v, against the aircraft, and meeting them in succession. At each meeting, the aircraft is removed from the system. If L is the total length of C, the time of exposure of each aircraft to collision with its neighbors is L/v (plus a small correction).

The third step in the approximation consists of a simplified description of the statistics of the proper motions of the aircraft reference points about their assigned positions, i.e., the equally spaced points, S units apart, regarded as fixed in the moving reference axis. If x_i is the distance, at the epoch t, of the i'th aircraft from its fixed reference position, a_i, x_i is a random quantity changing with time: a stochastic process – one, in fact, for each index i. Consider the pair of adjacent aircraft, $i = 1,2$, with $a_1 < a_2 = a_1 + S$. They will collide if, and only if, during the period L/v the random variables x_1 and x_2 acquire values such that

$$a_1 + x_1 > a_2 + x_2;$$

i.e., $$x_1 - x_2 > S.$$

If P_{12} is the probability of this event, we must know enough about the probabilistic features of the aircraft motion to calculate it, at least approximately. One assumption with some plausibility is to assume, first, that x_i and x_2 are probabilistically independent, and second, that each undergoes a diffusive change about its zero mean, as in a "random walk," symmetrical about zero. Then the distribution of each x_i is normal, with a standard deviation proportional to the square root of the elapsed time t. The same is therefore true of the difference. Thus, we write:

$$X_t = x_1 - x_2$$

and, assuming correct initial spacing, so that $X_0 = 0$, we must find the probability of the event

$$X_t > S \quad \underline{\text{for some } t \text{ between } 0 \text{ and } L/v}. \text{ (event 1)}$$

It will be observed at once that the stochastic quantity X_t is the type of limiting random walk known as the Wiener process (without drift), and that our problem is the classical one of an "absorptive barrier" at $X = S$: Will the time of first passage occur before L/v? This problem is solved in the standard texts such as reference 26. The solution is contained in Chapter 5, p. 221, formula (73) of reference 26, which gives the probability density g(t) for the time t of first passage. Replacing the drift constant μ and barrier constant <u>a</u> by 0 and S, the formula becomes

$$g(t) = \frac{S}{\sigma\sqrt{2\pi}} \cdot t^{-3/2} \cdot \exp(-S^2/2\sigma^2 t) \qquad (5\text{-}13)$$

where $\sigma\sqrt{t}$ is the standard deviation at time t.

The probability of collision [event (1)] is evidently

$$\underline{\text{prob.}}\ \underline{\text{collision}} \text{ of aircraft } 1 \ \underline{\text{and}}\ 2 = \int_0^{L/v} g(t)\, dt.$$

This integral is expressed in terms of the probability integral

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-x^2/2}\, dx \qquad (5\text{-}14)$$

by use of the change of variables

$$x = S/\sigma\sqrt{t}$$

$$\underline{\text{prob.}}\ \underline{\text{collis.}}\ 1\ \underline{\text{and}}\ 2 = 1 - 2\phi\left(\frac{S}{\sigma\sqrt{t_o}}\right) \qquad (5\text{-}15)$$

where $t_o = L/v$.

By taking the argument

$$\frac{S}{\sigma\sqrt{t_o}} = \frac{S}{\sigma}\sqrt{\frac{v}{L}} \qquad (5\text{-}16)$$

five or six units long, the probability of this event is acceptably small – so small, in fact, that there is no practical chance of any collision of any of the half dozen or so aircraft on the same path.

Note that, for a given spacing S and a standard deviation of a neighbor's distance per unit time σ, the chance of collision *decreases* with increasing speed v.

5.3.2 A Numerical Example. From reference 4 we learn that the arrival error (one standard deviation) of aircraft delivery at the landing strip is about 30 seconds. Assuming a landing speed of around 120 mph, this gives us a total standard deviation in the distance x_i from an aircraft to its assigned position, picked up during its total approach time $t_o = L/v$, of a quarter of a mile. Then the standard deviation of the distance X_t between aircraft assigned adjacent positions is this times $\sqrt{2}$, so that

$$\sigma\sqrt{t_o} = \sqrt{2/4} = 0.354 \text{ mile.}$$

If we wish the probability of collision of aircraft 1 and 2 not to exceed 10^{-6}, we must find for their minimum assigned spacing a value of S (in miles) equal to $s/\sqrt{2}$, where s satisfies

$$1 - 2\phi(s) = 10^{-6}.$$

In dealing with quantities as small as these, the standard tables are useless, and we have to replace $\phi(s)$ by its asymptotic expansion.[27] After obvious transformations, the above equation becomes

$$\exp(-s^2/2) \cdot \sqrt{2/\pi} \cdot \frac{1}{s}\left(1 - \frac{1}{s^2} + \frac{1.3}{s^4} - \frac{1.3.5}{s^3} + \ldots\right) = 10^{-6} \qquad (5\text{-}17)$$

As a first approximation, we drop all terms in the expansion after the first, and reduce the equation to

$$\exp(-s^2/2) = s\sqrt{\pi/2} \cdot 10^{-6} \qquad (5\text{-}18)$$

or, taking natural logarithms and setting $u = s^2$,

$$\log u = 12 \log 10 - \log \frac{\pi}{2} - u = 27.18 - u \qquad (5\text{-}19)$$

By a graphical method (observing the intersection of the curves $y = \log x$ with $y = 27.18 - x$) we find, to sufficient approximation, that $u = 24$, so that $s = 4.9$ and, finally, that the minimum spacing must be

$$S = 4.9 \cdot 0.354 = 1.73 \text{ miles.}$$

Next, let us consider the effect of *halving* this spacing -- which would *double* the rate of flow of aircraft to the landing field. Since this also halves s, we need only compute the previous asymptotic expression with $s = 2.44$ (in this case, we could use the error function tables directly). We obtain, instead of 10^{-6}, the probability of 0.013 (1.3 percent) of a collision!

Erwin[28] suggests that the capacity of an aircraft to respond quickly with changes of speed in the final landing process is severely limited. Therefore the random walk model may be a reasonably appropriate one in this instance. A further refinement would involve the effect of closing the control loop, that is, of purposely altering the speed or position of the aircraft on the basis of observations showing that it has strayed from its nominal position. This will result in a drift $x_1 - x_2$ which is no longer a Wiener process, the expected value of which grows in magnitude with time; but, under suitable assumptions, it can still be given a statistical characterization.

Analytical tools such as those used above are instructive for supplying orders of magnitude and for confirming our understanding of phenomena which are understood intuitively or empirically. Perhaps the greatest weakness of such methods is the assumptions that we make about the extremes of statistical distributions. When we make observations, they naturally concentrate around the zones of high probability density. Consequently, we can estimate means, medians, and variances, with reasonable assurance. However, analyses such as those above depend on the shape of the extremes of the distribution where the probability density is very low and where, therefore, we have essentially no observations. Evidence reported in the literature suggests that extreme excursions from nominal positions result from "blunders" and have a probability density much higher than one would infer from extrapolating a normal distribution with observed values of mean and variance.

5.4 INDIRECT MEASUREMENT OF RISK

5.4.1 Introduction. We must recall that the goal of this study is to develop terms, measures, and analytical tools for (a) the study of air traffic control capacity which are useful for the analysis of past experience, (b) the support of present decision-making, and (c) planning for the futt.re. For a measure or unit of safety to have any operational usefulness, it must fit into some scheme of analysis, decision-making, or planning. This in turn requires that the various measures and concepts correspond, directly or indirectly, to observable and measurable features of existing air transportation systems.

We have already seen that the penalty for lack of safety is almost entirely the loss of human life. Therefore, we should be able to make our principal safety concepts and terms correspond to observations about aircraft fatalities. The number of aircraft fatalities is quite considerable. In 1967, for example, there were 286 fatalities in accidents involving U.S. air carriers and 1 186 fatalities in accidents involving U.S. general aviation.[19] However, for the purposes of statistical analyses, these are not independent. The 286 air carrier fatalities occurred in only 12 accidents, and the 1 186 general aviation fatalities in 576 accidents. For the purposes of observation and rational measurement, we cannot regard multiple fatalities in a single accident as independent. If there is any degree of independence at all, we may be able to say that the respective fatal accidents are nearly independent of each other, but not the fatalities within one accident. Unfortunately, we need more than 12 events before we can draw significant conclusions about changes in accident rates. As shown in Appendix A – The Tyranny of Small Numbers – to make a valid inference that a desired 25 percent reduction is correlated with some change in condition, we must be able to observe something like 80 events in one condition and 60 in the other.

For a while, fatal accidents in general aviation may provide a useful foundation for decisions concerning safety, but the carrier fatalities are already below the useful level. Moreover, if we set an overall safety goal of the order of 1 fatal accident or less per 10 million hours of operation, then even in 1995, with a fivefold increase in air traffic, the goal would require reducing the total number of fatal aircraft accidents below 20 per year. We are faced with a possibility that our observations becomes less and less conclusive as we approach the goal. If there are several alternate routes for jointly increasing capacity and increasing safety, we would find it impossible to determine, by counting fatal accidents only, which route actually provides greater safety.

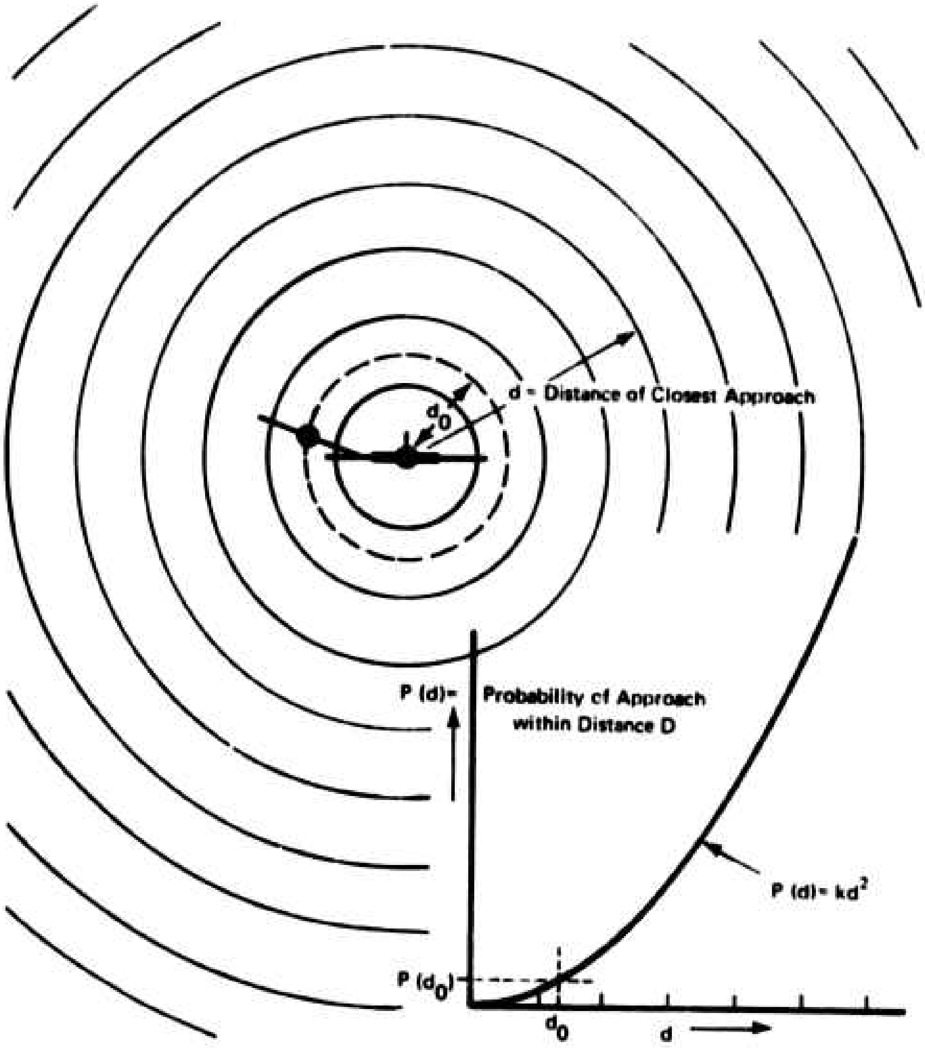
We must, therefore, look for indirect means of measuring safety. This will require the introduction of secondary units and auxiliary concepts to serve as a bridge between observations and the expected incidence of fatal accidents. The purpose of this section is to give an example of an indirect measurement related to safety. We shall show how an instrument capable of counting near misses out to a distance of 200 yards would be capable of collecting statistics about the probability of mid-air collisions on which meaningful decisions could be based in a time period as short as a year.

Imagine two aircraft with specified speed, aspect, and direction on a collision or near-collision course. Such a situation is unwanted, and is assumed to result from a mistake – either a big blunder or an accumulation of small equipment and human errors and failures, or some combination. There seems to be general agreement that near mid-air collisions are far more often the result of gross blunders than they are of accumulation of small errors.[5,12] Since such a situation is not planned, it is reasonable to assume an element of randomness in the encounter.

Assume this model of randomness: Imagine that, if the situation were repeated, each aircraft might fly at the same course and speed, but be displaced so that the probability distribution of their paths is uniform in the plane perpendicular to its motion for distances of many hundred feet to all sides. Under this assumption (see Figure 5-2), the probability that the two aircraft will approach to within a distance d is proportional to d^2, say kd^2. This is graphed in the lower right portion of the figure, as is the fact that at some minimum distance d_0 (dependent on factors such as aspect and relative course) the two aircraft will collide and, with high probability, crash.

Actually, the assumption of uniform probability distribution is too restrictive. It is quite sufficient that the probability distribution of the trajectory of one aircraft be a harmonic function* in a plane perpendicular to the line of relative motion. Because of the strong averaging implied in both this step and subsequent steps, the results are quite insensitive to the shape of this probability distribution, even if it fails to be harmonic or constant.

*In the narrow mathematical sense.



d = Distance of Closest Approach
d_0
P (d)= Probability of Approach within Distance D
P (d)= kd^2
P (d_0)
d_0
d

FIGURE 5-2 RANDOMNESS MODEL SHOWING AIRCRAFT FLYING ON SAME COURSE AT SAME SPEED, BUT DISPLACED SO THAT PROBABILITY DISTRIBUTION OF THEIR PATHS IS UNIFORM

(Graph at lower right shows that at some minimum distance, d_0, the two aircraft will collide and, with high probability, crash)

If we assume some degree of statistical independence among the various occasions under which these near encounters take place, then the number of approaches to within a distance d has a Poisson distribution, the expected value of which is proportional both to d^2 and time of exposure.

Now let us integrate this picture over all aircraft pairs, over all courses, speeds, aspects, and other variables, weighting each situation in proportion to its actual likelihood of occurrence. One objection might be that we do not know the values of these likelihoods or, for that matter, any of the other parameters. It does not matter. The proportionality to d^2 is maintained, $P(d) = \bar{k}\, d^2$, but with an averaged value of the constant of proportionality $\bar{k}$. Also, the collision probability is averaged and is equal to the probability of approach within an averaged collision distance $\bar{d}_0$. As a first approximation, $\bar{d}_0$ is the radius of a circle, the area of which is four times the average area of the aircraft's silhouette; hence a few score feet. Reference 24 estimates d_0 at about 50 feet; and the average collision cross section is estimated by a method which leads to estimates of d_0 of 15.6, 28.7, and 53 feet for encounters between two light aircraft, one light and one heavy, or two heavy aircraft. There is no conceptual problem in making a precise estimate of this number.

Note now that the probability of an encounter with an approach distance of 500 feet, for example, is 100 times as great as the probability of collision. This is the central idea of this indirect measurement.

Suppose we could measure accurately the distance of closest approach out to 700 or 1000 feet, for instance. Also, suppose we could instrument 20 percent of the air carrier fleet, about 1 million hours of operation per year, and tabulate the number of close approaches in a year as a function of distance. For the sake of a concrete illustration, let us postulate the data displayed in Table 5-1. We have accumulated data about 1 million hours of operation in each of two regimes – Regimes I and II. The number of events with miss distance d_0 is displayed in the table as a function of the

TABLE 5-1

HYPOTHETICAL COUNT OF NEAR ENCOUNTERS

		Number of Events with Miss Distance Less Than			
	Total Hours	200 ft	300 ft	500 ft	700 ft
Regime I	1,000,000	6	11	24	50
Regime II	1,000,000	2	3	12	20

distance. These results are also graphically displayed in Figure 5-3. The scales are logarithmic, so the square law dependence plots are straight lines. In our hypothetical case, there were 50 approaches to within 700 feet during 1 year of operation under one regime and 20 approaches to within the same distance in the following year under a new regime. The difference between 50 events and 20 events is statistically significant at a confidence level of better than 0.001. This is best illustrated in Figure 5-4, where the number of close approach events is plotted against exposure on binomial probability paper, which is designed to normalize the distribution and regularize its standard deviation. The area of the 700-foot radius circle is $1.54 \cdot 10^6$ square feet. With a total exposure of 2 million hours, the normalized exposure is $3.08 \cdot 10^{12}$ square-foot-hours, during which 70 events took place. For a test of statistical significance, we assumed the null hypothesis: the rate of events is 70 events in $3.08 \cdot 10^{12}$ units of exposure, and the two regimes are not significantly different. Following the graphical methods of Mosteller and Tukey,[29] it is easy to show that the difference between 50 events and 20 events (at 700 feet) is highly significant, but the difference between 24 events and 12 events (at 500 feet) is barely significant at a confidence level of around 10 percent.

Relying on the quadratic relation between probability and miss distance, we can now infer that the latter regime is safer than the former, even though the inferred collision probabilities are much less than 1 per 1 million hours of operation.

We can also test the quadratic relation between probability and distance by counting numbers of approaches to within closer distances – 500, 300, 200 feet, and so forth. Standard statistical tests of significance will tell whether the observed numbers are consistent with the assumed probability distribution. We are suspicious of the data in the table, in fact, because they are somewhat too regular to be convincing.

However we use these data, the measurements must be fairly accurate and the sample of close approaches measured must be unbiased with respect to distance of closest approach. Near-miss reports, as we presently know them, are not adequate for this purpose, although they do have many uses. An unbiased measurement error with a standard deviation of 95.6 feet, as reported in reference 24, would be smoothed considerably in a cumulative plot of 20 or 50 events. However, a systematic variation of 150 feet, the magnitude predicted by extrapolating Figure 9 of reference 24 to a

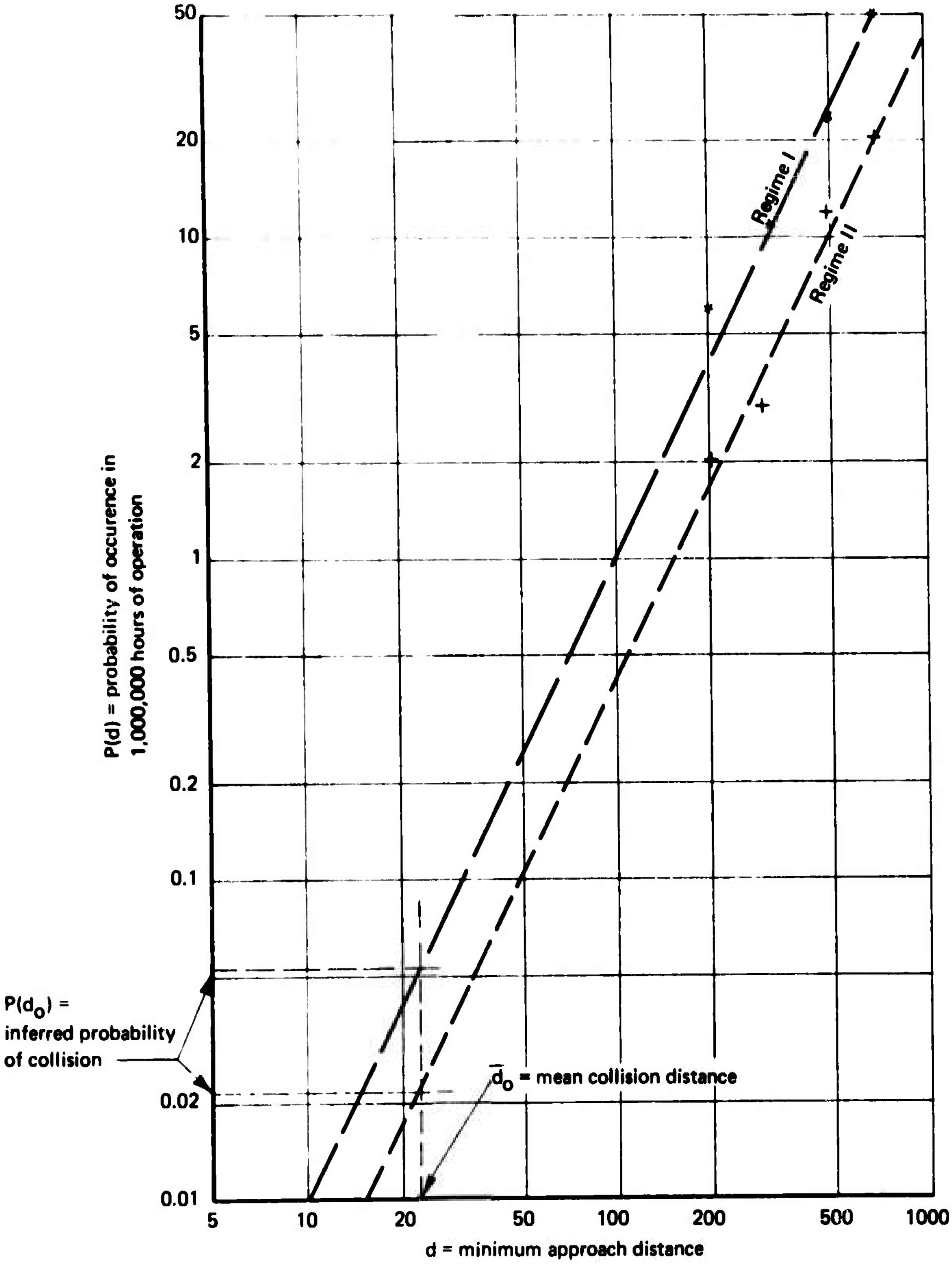


FIGURE 5-3 FREQUENCY OF NEAR ENCOUNTERS AND PROBABILITY OF COLLISION

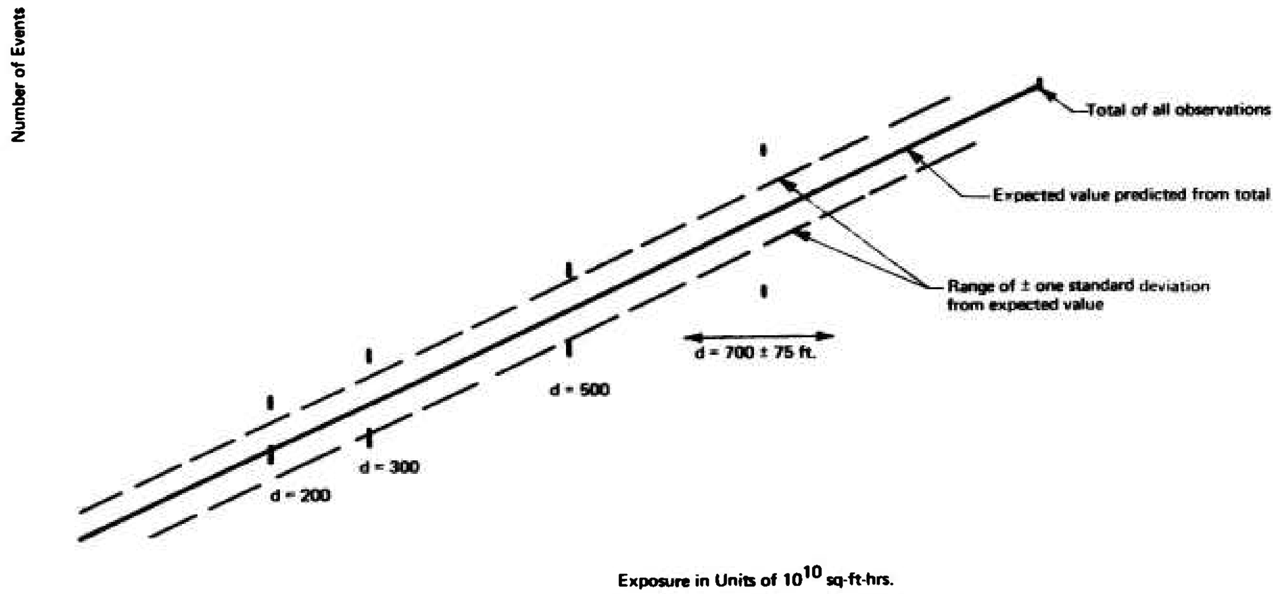


FIGURE 5-4 NUMBER OF CLOSE APPROACHES AS A FUNCTION OF EXPOSURE (Hypothetical Data of Figure 5-1)

distance of 700 feet, would be intolerable. A line segment is plotted on Figure 5-4 showing an uncertainty of ±75 feet in the units of the abscissa. Such an uncertainty, if systematic, would wipe out the significance of these numerical results.

Referring again to Figure 5-3, let us examine the inferred probability of collision. Using the average mean collision distance estimate of reference 24, approximately 23 feet, we can estimate the probability of collision as approximately 0.06 per million hours of operation with Regime I and 0.024 per million hours of operation with Regime II. These are both less than one-tenth of the recently observed rate of fatal accidents from all causes in our safest large category, domestic scheduled air passenger flight, where the rate of fatal accidents has hovered around 1 per 1 million hours of operation for a number of years. They are also below the level which Starr[20] finds where people behave as though a voluntary risk were negligible.

5.4.2 Conclusions. The probability of an encounter between two aircraft at a distance of several hundred feet is two or three orders of magnitude larger than the probability of mid-air collision. A count of such events among aircraft with an aggregate operating time of 1 million hours can produce a data base from which the relative safety of one regime over another could be statistically validated, even if both have an expected value much lower than the presently encountered probability of mid-air collision involving domestic scheduled carrier aircraft.

REFERENCES

1. Report of the Department of Transportation Air Traffic Control Advisory Committee, Volumes I and II, December 1969.

2. Marner, G.R., *Conceptual Questions in Air Traffic Control Design*, Collins Radio Company; paper presented before the Institute of Navigation National Air Meeting on Air Traffic Control in the 1970s, April 15, 1970.

3. Goldman, A.J., *Analysis of a Capacity Concept for Runway and Final-Approach Path Air Space*, National Bureau of Standards; paper presented at the Institute of Navigation National Air Meeting on Air Traffic Control in the 1970s, April 15, 1970.

4. Astholz, P.T., Scheftel, D.J., and Harris, R.M., *Increasing Runway Capacity*, Proc. IEEE, Vol. 58, No. 3, March 1970, pp. 300-306; also *Airport Design Considerations*, App. B-7 of the ATCAC Report (ref. 1).

5. Steinberg, H.A., *Collision and Missed Approach Risks in High Capacity Airport Operations*, Proc. IEEE, Vol. 58, No. 3, March 1970, pp. 313-321; also *A Safety Model for Evaluating Risk Involved in Airport Landing Operation*, App. B-2 of the ATCAC Report (ref. 1).

6. *FAA Statistical Handbook of Aviation*, Chapter 10, Department of Transportation, 1967.

7. Fromm, G., *Economic Criteria for Federal Aviation Agency Expenditures*, prepared for the FAA under Contract No. FAA BRD-355, United Research Incorporated, June 1962.

8. Holt, J.M., and Marner, G.R., *Separation Theory in Air Traffic Control System Design*, Proc. IEEE, Vol. 58, No. 3, March 1970, pp. 369-376; also *Separation Hazard Criteria*, App. C-4 of the ATCAC Report (ref. 1).

9. Hagerott, R.E., and Weiss, W.J., *Aircraft Separation Minima, Standards, and Criteria and Related National Air Space System Component Variables*, draft prepared by the Group 3 ATC Advisory Committee, February 7, 1969.

10. Ratcliffe, S., and Worcestershire, M., Royal Radar Establishment, paper prepared for the United Kingdom Symposium, "Electronics for Civil Aviation," 1969.

11. Working Paper, *Reference Material for Air Traffic Control Separation*, compiled for the Air Traffic Control Advisory Committee, Group 3, January 1969.

12. Reich, P.G., *A Theory of Safe Separation Standards for Air Traffic Control*, Royal Aircraft Establishment, Technical Report No. 64041, November 1964.

13. Reich, P.G., *An Analysis of Planned Aircraft Proximity and Its Relation to Collision Risk*, with Special Attention to the North Atlantic Region, 1965-71, Royal Aircraft Establishment, Technical Report No. 64042, November 1964.

14. Marks, B. L., *Air Traffic Separation Standards and Collision Risk*, Royal Aircraft Establishment Technical Note No. Math 91, February 1963.

15. Raisbeck, G., *Study of Air Traffic Control System Capacity Measurement Methodology: A Six-Month Progress Report*, Arthur D. Little, Inc., March 1970.

16. *Interview Guide and Questionnaire on Factors Affecting Air Traffic Separation*, Working Group 3, Air Traffic Control Advisory Committee, January 1969.

17. McFadden, Col. J.G., *Impact of Helicopter Operations in ATC in the 1970s*, U.S. Army, AASC, presented at the Institute of Navigation National Air Meeting on Air Traffic Control in the 1970s, April 15, 1970.

18. Schelling, T.C., *The Life You Save May Be Your Own*, in Problems in Public Expenditure Analysis, Samuel B. Chase, Jr., ed., The Brookings Institution, Washington, D.C., 1968.

19. *FAA Statistical Handbook of Aviation*, Department of Transportation, 1968.

20. Starr, C., *Social Benefit Versus Technological Risk*, Science, Vol. 165, No. 3899, September 19, 1969, pp. 1232-1238.

21. Graham, W., and Orr, R.H., *Terminal Air Traffic Flow and Collision Exposure*, Proc. IEEE, Vol. 58, No. 3, March 1970, pp. 328-336; also *Terminal Air Traffic Model with Near Mid-Air Collision and Mid-Air Collision Comparisons*, DOT Air Traffic Advisory Committee Report, Vol. 2, Appendix C-3 (ref. 1).

22. Alexander, B., *Aircraft Density and Mid-Air Collision*, Proc. IEEE, Vol. 58, No. 3, March 1970, pp. 377-380.

23. Koopman, B. O., *Search and Screening*, Operations Evaluation Group, Summary Technical Report II-B of Division 6, Office of Scientific Research and Development, National Defense Research Committee, Washington, D.C., 1946.

24. Graham, W., and Orr, R.H., *Separation of Air Traffic by Visual Means: An Estimate of the Effectiveness of the See-and-Avoid Doctrine*, Proc. IEEE, Vol. 58, No. 3, March 1970, pp. 337-360.

25. *Near Mid-Air Collision Report of 1968*, Air Traffic and Flight Standards Technical Report, Advance Report, July 1969.

26. Cox, D.R., and Miller, H.D., *The Theory of Stochastic Processes* (New York, John Wiley & Sons, 1965).

27. Peirce, B.O., *A Short Table of Integrals* (Ginn & Co., Boston, 1929) p. 120.

28. Erwin, R. L. Jr., *Influence of Flight Dynamics on Terminal Sequencing and Approach Control,* Report of the Department of Transportation Air Traffic Control Advisory Committee, Vol. 2, December 1969, Appendix B-5.

29. Mosteller, F., and Tukey, J.W., *The Uses and Usefulness of Binomial Probability Paper,* J. Am. Stat. Assn., Vol. 44, June 1949, pp. 174-212.

6. TIME-DEPENDENT QUEUES APPLIED TO THE STUDY OF CAPACITY

6.1 INTRODUCTION

One of the most powerful methods of answering difficult questions concerning ATC capacity, delay, and related matters of air transportation is the *computer-implemented analytical study of time-dependent queues.* The practical applications of queuing theory to operational problems go back to the beginning of this century, and have resulted in a massive body of literature with which the ADL team working on this study is familiar. Surprisingly enough most of it is inapplicable to the ATC problems under study in this report, since most of the published work ignores the *time dependence* of inputs and other conditions in the ATC problems, and is for the most part confined to steady state studies.*

While these papers supply a certain amount of general background, none of them deals with formulations directly relevant to the present study, nor do they give explicit solutions to their problems. Moreover, they fail to treat such important matters as periodicity, and the like.

Another mode of attack bypasses most of the analytical reasoning and applies computer simulation. Probably the best known publication in this category, "Nonstationary Queuing Probabilities for Landing Congestion of Aircraft."[1] has all the limitations of a *simulation* as opposed to a *computation.*

Because of the insufficiency of methods in the available literature, for our ATC problems the ADL team has had to develop new ones. These methods and a number of informative results of applying them in sample cases are set forth in this chapter.

In Section 6.2 of this chapter is included an analysis of the situation arising in air traffic when planned schedules and random variations are mixed, resulting in the various queues, delays, and other influences which have an effect on capacity. Then results of applying our analytical tools are given and interpreted with the aid of graphs in Section 6.3, although the full force of these tools is only sampled rather than applied to every possible circumstance. The mathematical details are set forth fully in Section 6.4 in the case of the single queue and one runway, and in Section 6.5 in the far more complicated case of two queues (landing and takeoff) using one

*See Addendum to this chapter.

Preceding page blank

runway. Delay times are discussed in Section 6.6. The extension of the analytical treatment to multiple runways is the natural next step, but its intensive examination is a subject for further work and is not included in this report.

With the arrangement of material in this chapter, the general reader can limit himself to Sections 6.2 and 6.3, which form a self-contained presentation. The highly technical mathematical material, while forming the basis of the presentation, is not necessary to its understanding. Some of Section 6.6 is also only of general interest.

Before ending these introductory remarks, a comment regarding our technical approach is well in order: We have used *computer-implemented analytical methods* rather than *computer (Monte Carlo) simulations* in an area of work in which the latter method is more often used in this country. More precisely, our procedure consistsof the three following steps:

1. A quantitative description of the system studied and the approximate assumptions;

2. An embodiment of assumptions concerning the evolution of the system (its states and their probabilities) in precise mathematical statements (equations); and

3. A study of the general properties of the solution and its numerical evaluation by appropriate mathematical methods and the use of computers.

The method of computer simulation, on the other hand, carries out step 1, but replaces steps 2 and 3 by the process of step-by-step changes in the state of the computer, the rules for these changes being programmed into it, and intended to correspond to the changes in the state of the actual system under study.

When the situation considered is completely *deterministic*, the method of simulation is simply a special instance of the use of the computer for the numerical solution of the equations that determine the problem (equations which are not necessarily written out explicitly). When, on the other hand, the situation involves *random* and requires probabilities, expected values, standard deviations, and the like, the change of state in the simulation has to follow a Monte Carlo process, usually based on the use of tables of random numbers. In this case, the advantage of the analytical methods we are using over the simulations is, first, in the *greater efficiency* *in the use of the computers for obtaining a sufficient level of reliability of the numerical answers;* and, second, – and this applies also in the deterministic cases – in the *incapability of computer simulations to establish general properties* of the ATC system. Thus it is important for practical reasons to show that if the inputs of the system have a diurnal periodicity (a 24-hour recurrence), the same will be true of some of the solutions, whereas others will merely approach a periodically varying solution. It is also of great practical importance to find whether the solutions are stable or unstable, and to establish estimates of the amounts of fluctuation from stable situations: unacceptable delays and capacity overloads may be predictable on such bases. The best that can be done by simulation is to obtain numbers that *suggest* such effects.*

6.2 ANALYSIS OF THE PROBLEM

In Chapter 3, the first definitions of the terms and measures related to capacity, demand, and delay were based on the picture of streams of aircraft moving in a predictable manner, as in the flow of a material substance, such as a fluid. Toward the end of Section 3.3.4, Stochastic Models, the need for a more realistic picture that, in contrast to the deterministic model, recognized the essential element of *random* in most air traffic operations was explained. The sources of this random were also explained, and their practical consequences indicated in general qualitative terms. Sections 3.4 and 3.5 of Chapter 3 again indicated the eventual need to take these random elements into account in the basic terms and measures.

Using, as in Chapter 3, the single air terminal as a basic illustration of the situations of concern to ATC, let us first consider the *arrivals* of aircraft intending to land. In what numbers will these aircraft come within a conventionally established distance of the air terminal – beyond which they are uncoordinated (except through a time table), and within which they come under ATC direction? Because of the random element in the arrivals, all that can be stated is in terms of the *probabilities* of various numbers (0, 1, 2, etc.) of arrivals during a given interval of time, as between 8 a.m. and 8:05 a.m.; in more general terms, during a given (short) interval of time $(t, t + \Delta t)$ (in the example, $t = 8$ a.m. and $\Delta t = 5$ minutes).

*Statements are frequently made implying that computer simulation can deal effectively with a broader class of problem than analytical methods. If this is a statement of a principle, it cannot be accepted. Computer simulation which is not based on quantitative reasoning (mathematics) gives, at best, an intuitive suggestion, but can prove nothing reliably, whereas, given the quantitative reasoning, an analytical formulation is always possible in principle – with a degree of ease depending on the luck and ability of the worker

The following assumptions are usually made as natural approximations to the complex reality:

- The probability of non-arrival tends to unity as Δt is shorter and shorter; that of a single arrival is more and more nearly proportiona! to Δt (and can be written as $\lambda\Delta t$); that of two or more simultaneous arrivals is much smaller, and of the order of $(\Delta t)^2$.

- Different arrivals are *independent* events. This means, for example, that even when it is known that an aircraft reaches LaGuardia from Chicago during $(t, t + \Delta t)$, this does not change the probability of one reaching it from Boston during the same period. This is a statement regarding "conditional probabilities."

A necessary consequence of these assumptions, easily derived by probability reasoning, is that the probability of exactly k arrivals during the short period $(t, t + \Delta t)$ is

$$e^{-\lambda \Delta t} \ (\lambda\Delta t)^k/k! \ (k! = 1.2 \cdot \ldots \cdot k; 0! = 1).$$

This is the well-known *Poisson Law of Occurrences.* It will be assumed in the present study as a useful approximation; it will be further assumed that

- *the arrival rate parameter* λ *introduced above may depend on the time of day* t*:* $\lambda = \lambda(t)$, and that the degree of variation in this quantity, while considerable during the course of the day (it may be about zero in the early morning hours and rise to high values in high traffic periods) varies slowly enough to make the Poisson Law sufficiently accurate. In Section 6.4, $\lambda(t)$ is assumed to be a periodic function of t with a period of 24 hours.

When the aircraft arrive (as described above) at a greater rate than they can be "serviced" – i.e., are allowed to land – they are obliged to enter a holding pattern. The aircraft in this pattern, as well as those in the landing pattern, are to be regarded as in a *queue* or waiting line. Here a point of view will be adopted that is intended to reflect the limited number of aircraft that can be allowed in such a queue – because of the limited volume in which aircraft can be stacked, the limited number that can be kept under ATC, the limited endurance of the aircraft, and so forth. If the limit allowed in the queue is denoted by m,* then all arrivals which would make the queue increase beyond m are *diverted* – to other terminals, or held at their point of origin. Part of our study of the queues will deal with the probable numbers

*m = 25 in the example of Section 6.3, but it may be higher or lower, depending on the terminal's facilities, weather conditions, etc.

that may have to be diverted. Another part of this study (Section 6.6) will consider the probabilities of the various *delays* caused by waiting in the queue.

Two other matters must be settled, in addition to the Poisson Law of arrivals and the limited number allowed in the queue, if we are to be able to get hold of the probabilities of the various possibilities at the terminal – indeed, if these probabilities are to be *determinate*: we must have a law of landing and of the time taken for this landing. It might at first be supposed that this is a perfectly regular and predictable process: if, at a given time, there are k aircraft in the waiting line ahead of the one we are in, and it takes time T for each to land, then we will have to wait a length of time kT before landing in the further time T. Such an extreme of regularity is unrealistic. There are too many chance departures from it, both in the time taken to land, and in the opportunities of gaining access to the landing strip – which is also used by other aircraft for takeoff.

The law of landing must evidently be probabilistic, and it must reflect the regular (deterministic) and random aspects of the problem. Under such conditions, the compromise which is mathematically simplest is to assume a Poisson Law, based on a parameter μ which may be independent of the time of day, but possibly dependent on the state of the ground waiting line for takeoff and also (under certain queue disciplines) dependent on the landing queue. For a single runway, such an assumption may be unrealistic, because it allows landings to come much closer together than is actually allowable. However, we should not exclude it on that account, any more than we reject mass points and weightless strings in the study of mechanics: Such an assumption may be a perfectly adequate basis for the representation of variables other than the distribution of interarrival intervals. If the assumption gives results that are in reasonable accord with what is observed, and if they are also not very different from the results of assumptions of regular behavior, it has passed its first test and can be regarded as leading to a promising model. This assumption, in its various forms, will be used in the later sections of this chapter.

With no additional complexity, the service rate may be made a function of the time of day also. There are well-known methods (which unfortunately increase the mathematical complexity considerably) for dealing directly with the servicing times, that is, the time taken to land and the time taken to take off. We can attribute both regular and random factors to both of these.

The Poisson assumption above is equivalent to the assumption that the servicing times have exponential distributions. Results – not very different – can be obtained by assuming that it has a given constant value, or that it has some other statistical distribution. In the remaining parts of this chapter, the exponential distribution of service times, consistent with the Poisson distribution of moments of initiation of service, are adopted because they are the simplest to deal with mathematically.

In Section 6.4 a single queue – the "landing queue" – is studied by setting up the differential equations governing the $m + 1$ probabilities $P_n(t)$ that, at time t, there are just n aircraft in the queue $(n = 0, 1, 2, \ldots, m)$. These turn out to be relatively easy to deal with, being of a familiar form (first order, linear, and homogeneous). For any arrival parameter $\lambda(t)$, given by a graph or a table, an altogether reasonable computer program leads to the solution: the values of each $P_n(t)$ for all times, as well as expected (average) number as functions of the time, standard deviations, and *times* taken from arrival to landing. The results of this process are given and discussed with the aid of graphs in Section 6.3

In Section 6.4 it is also proved mathematically that when $\lambda(t)$ is a periodic function with a 24-hour period, the same will be true of one solution, whereas other solutions will not have this property but will approach the one that does as time goes on.

In Section 6.5 the much more difficult – but feasible – problems arise when both the landing and the take-off queues, viewed as interacting together, are considered under various queue disciplines. While programs are given for numerical solution of the equations by computers, the results are not analyzed in this report.

It may be observed that queues may occur at many other points during a flight, the results giving rise to queues in tandem: the output of one being the input of the next, and so forth. Further, multiple runways give rise to other situations, e.g., parallel queues.

We close this section with a glossary of terms, serving both as a summation of what has been discussed at length, and as a reference in the mathematical discussion given in later sections.

GLOSSARY OF TERMS

QUEUE – A queue is formed when customers (the word customer is used in a technical sense; in practice it might, for example, have to be equated with 'aircraft') arrive at a service station (offering certain facilities) and demand service. A queue at any point in time will consist of customers waiting for service as well as those receiving service. A waiting line will consist of customers actually waiting to be served. A queuing system is completely described by (1) the input, (2) the queue discipline, and (3) the service mechanism.

INPUT -- An input describes the way customers arrive and join the system. The number of customers may be finite or infinite, and they may arrive individually or in groups. The rule governing arrivals may be deterministic or a stochastic process. The simplest hypothesis about the input is one which states that the customers arrive at 'random' (i.e., in a Poisson process), the number of arrivals in time t being a Poisson variable of expectation λt. The time interval u between two consecutive arrivals will then have the exponential density:

$$\lambda e^{-\lambda t} \ \Delta t.$$

The distribution of the intervals between arrivals is called the *inter arrival distribution* and the total input with its specification is called the *arrival process.*

A QUEUE DISCIPLINE - A queue discipline is the rule by which customers are chosen for service by the servers, e.g., first-come/first-served, last-come/first-served, random service, and so forth.

SERVICE MECHANISM – A service mechanism is the arrangement for serving customers. In general, there are N servers where $N \geqslant 1$. Usually all servers are available, but there are situations where one or more of them will be absent from the system at certain times. If $N < \infty$, the servers attend the customer in a specified order, e.g., in one case, the first of the N servers to be free attends the customer at the top of the queue. The time t which elapses while a particular customer is being served is called his *service time* and the distribution of t, the *service time distribution.*

WAITING TIME – Waiting time is the time spent by a customer in the waiting line before commencement of his service, i.e., if a customer arrives at time n and enters service at time y, then his waiting time is y-n. The distribution of y-n is the *waiting time distribution.*

STEADY STATE DISTRIBUTION -- If as $t \to \infty$, the distribution of various quantities converges to distributions, independent of the initial conditions, these latter distributions are called the steady-state distributions. In general, when a steady-state distribution exists, the queue can be started according to these rules and the various distributions will then be invariant in time, e.g., $Pn(t) = P[n$ customers in the queue at time $t]$, i.e., the probability that there are n customers in the queue at time t will be independent of t.

BUSY PERIOD -- A busy period begins when a customer enters the system and there are no previous customers in the system, and it ends the next time the system is empty.

PRIORITY QUEUES – Customers with different priorities arrive as inputs of the same or different arrival distributions, wait to be served on a first-come/first-served basis within each priority, and are served by one or more servers. A low priority customer may (preemptive service) or may not (non-preemptive) be ejected back into the line when a higher priority item enters the system.

BULK QUEUES – A bulk queuingsystem results when either the arrivals or the service, or both, occur in groups (or batches); e.g., several people may go to a restaurant together and obtain service as a group, and so forth.

6.3 GRAPHICAL PRESENTATION OF SAMPLE RESULTS

Section 6.2 covers the background of assumptions regarding the laws of arrival, departure, and time to land. In Sections 6.4 and 6.5, these assumptions have been used to set up the basic equations of evolution of the state of affairs about a one-runway airport, considering either the single queue (Section 6.4) or the air and ground queue combination. The resulting differential equations, having coefficients given by graphs or numerical tables rather than by formulas, are appropriately solved by computers.

The object of this section is to illustrate the practical results that can be obtained from the above process. For this purpose, the single (landing)-queue situation of Section 6.4 has been chosen; with moderate adaptation its results have wider application. To complete the illustration, two air terminals (A and B) are examined. Terminal A has its *arrival rate* $\lambda(t)$ coincident with what was actually shown* by statistical observation at J.F. Kennedy for one

*Private communication from the FAA.

month in 1968, while similar data observed for LaGuardia were used for Terminal B. As for the quantity μ, the service rate (number landing per unit time) is given three constant values – 45, 55, or 75 aircraft per hour – chosen in Terminal B to correspond very roughly with those of LaGuardia, and at the same time – by the use of three different figures – to bring out the degree of sensitivity of the results to this parameter. In the case of Terminal A, mere orders of magnitude are intended to be reasonable, but there is little detailed resemblance to J.F. Kennedy which may have several simultaneously used runways. After these three constant values of μ are chosen, a fourth non-constant case having a sharp dip for a moderate length of time, is examined. This is intended to explore the effects of a brief weather upset or other misadventure that temporarily slows down the landing rate.

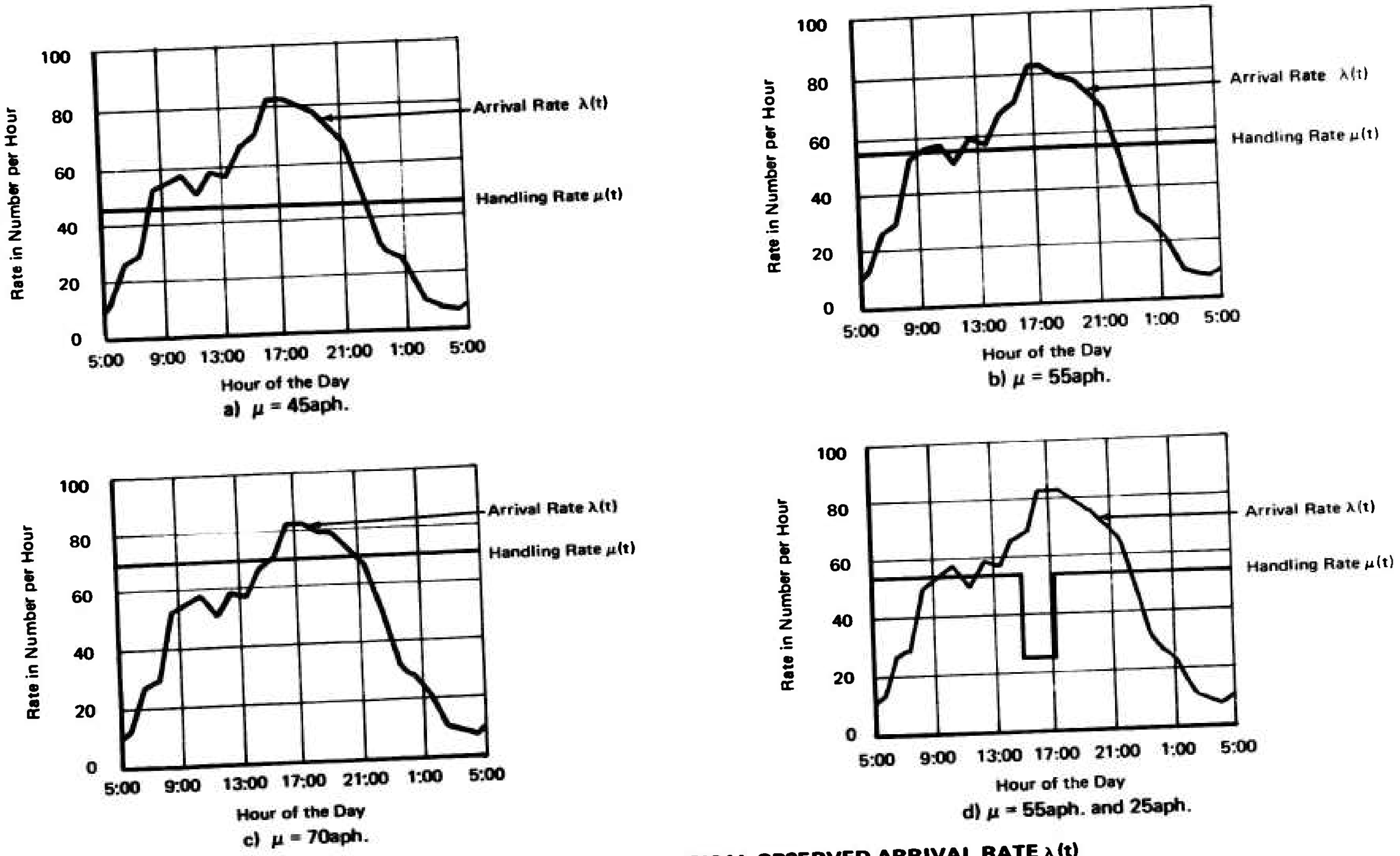
This combination of the two arrival rate profiles and four services rates gives eight sets of graphs* (Figures 6-1 through 6-10). In each set, the functions of chief interest are graphed in pairs: probability of zero or of maximum allowed numbers in the queue (i.e., $P_0(t)$ and $P_m(t)$); expected number $E(N)$ or $\bar{N}$ and standard deviation; waiting times (except with the non-constant service rate); and numbers turned away (not necessarily physically; they may be diverted or held on the ground, etc.); and other functions, as indicated.

It should be clear from these graphs that the somewhat abstractly formulated mathematical tools of analysis do, in fact, provide absolutely concrete results, and since they are obtained by calculation from stated assumptions rather than by Monte Carlo simulation, their numerical precision and reliability can be subjected to full scrutiny.

Another point that the graphs should emphasize is the strong time dependence of the whole situation: nothing like a "steady state" is anywhere to be seen. This is why we have had to develop our own analytical tools, rather than using existing ones.

It may be noted that the input values of the arrival rate function $\lambda(t)$ and the service rate μ, which "drive the results," may be regarded as periodic functions with a 24-hour period. To show this graphically, we would draw a time scale of length many times 24 hours, fill in the stipulated curves for $\lambda(t)$ and μ for one 24-hour period, and then repeat the same curves displaced ±24 hours, ±48 hours, etc., to produce a periodic curve. When the output curves

*Figures 6-1 through 6-5 include 4 sets (one arrival rate and four service rates) and Figures 6-6 through 6-10 include the other four sets (one arrival rate and four service rates).



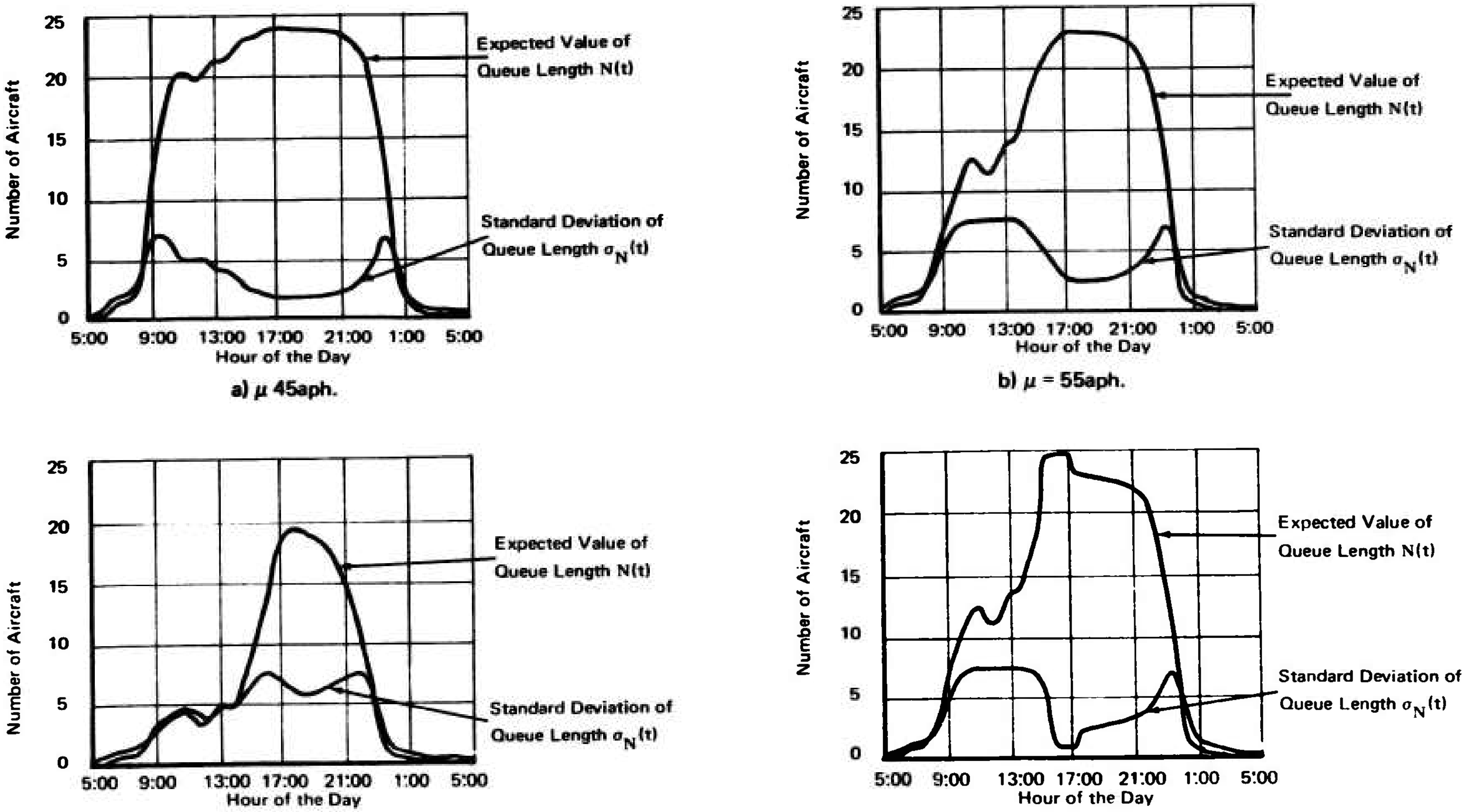
a) μ = 45aph.

b) μ = 55aph.

c) μ = 70aph.

d) μ = 55aph. and 25aph.

FIGURE 6-1 TERMINAL A. A TYPICAL OBSERVED ARRIVAL RATE $\lambda(t)$ AND FOUR ASSUMED HANDLING RATES $\mu(t)$



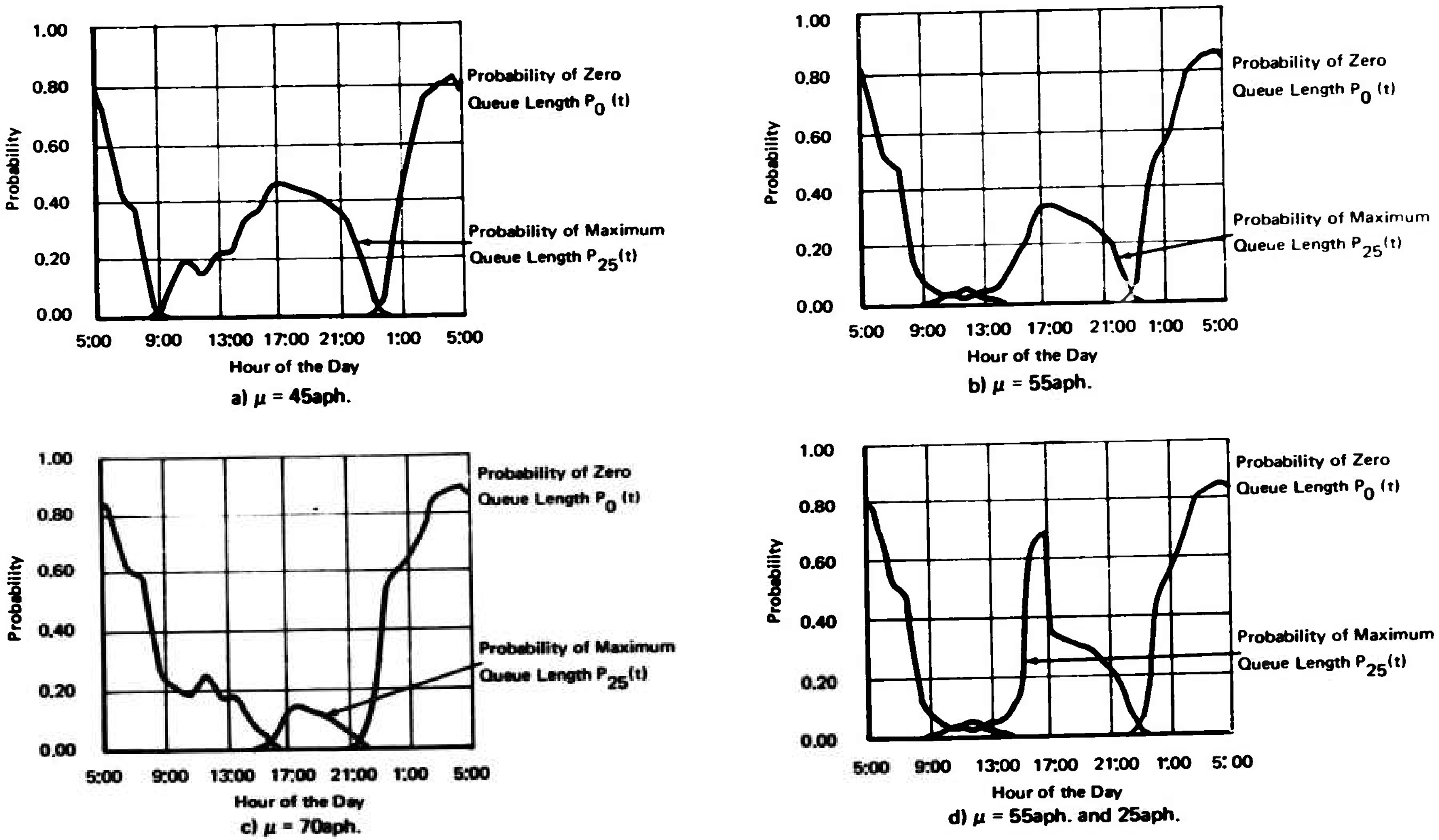
a) μ 45aph.

b) μ = 55aph.

c) μ 70aph.

d) μ = 55aph. and 25aph.

FIGURE 6-2 TERMINAL A. EXPECTED VALUE N(t) AND STANDARD DEVIATION σ_N (t)OF QUEUE LENGTH; ARRIVAL AND HANDLING RATES AS IN FIGURE 6-1



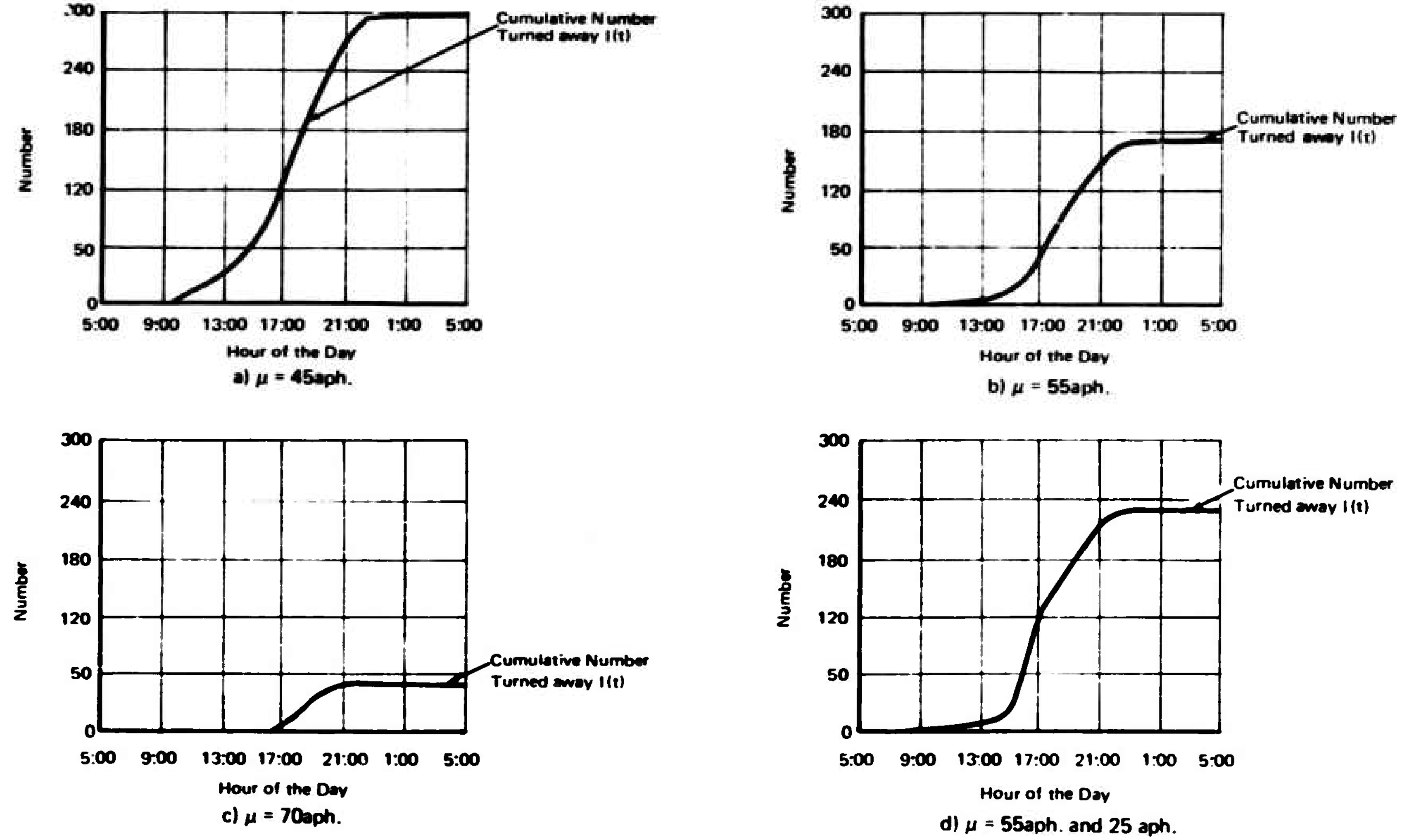
a) μ = 45aph.

b) μ = 55aph.

c) μ = 70aph.

d) μ = 55aph. and 25aph.

FIGURE 6-3 TERMINAL A. PROBABILITY OF MAXIMUM QUEUE LENGTH P_{25}(t) AND ZERO QUEUE LENGTHP_0 (t); ARRIVAL AND HANDLING RATES AS IN FIGURE 6-1



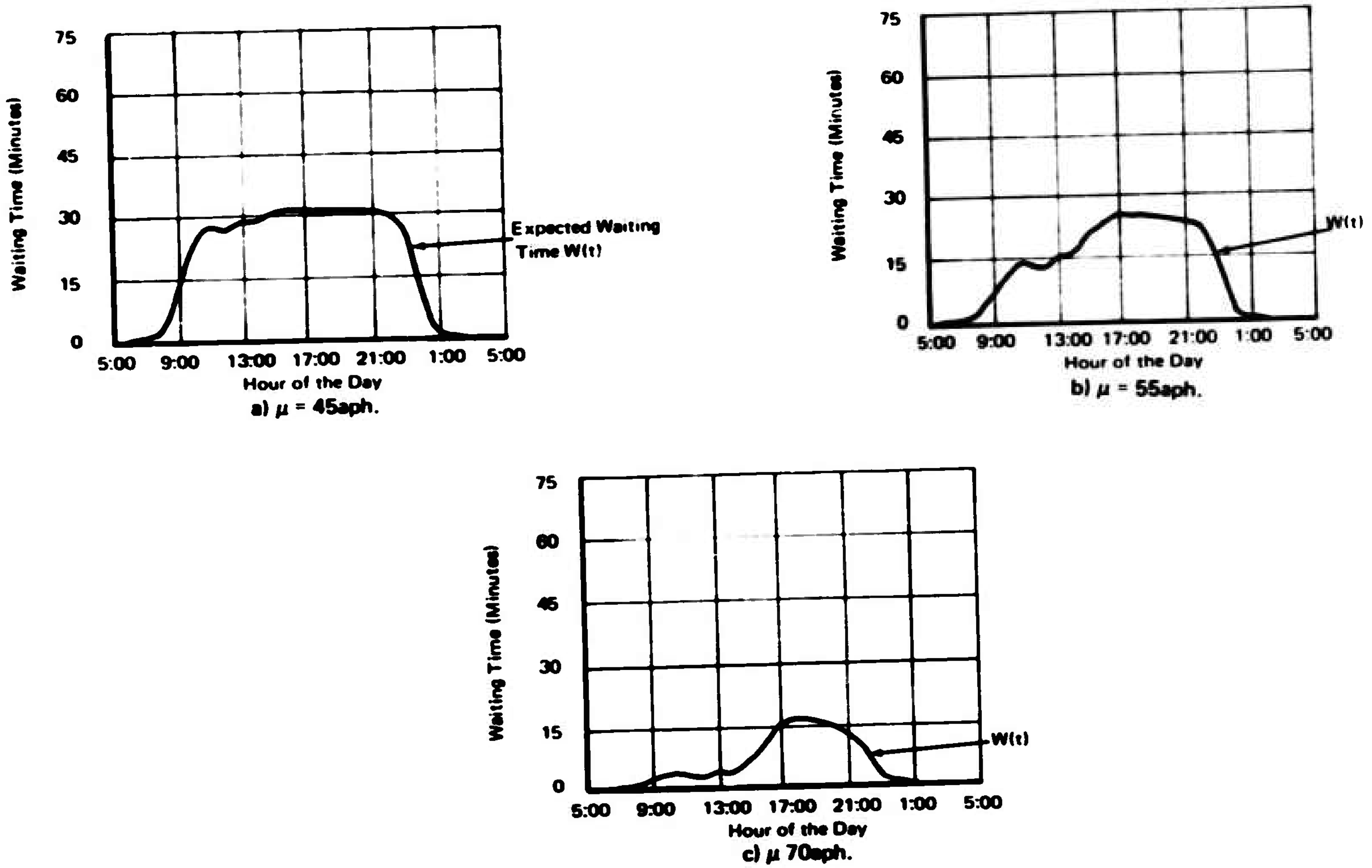
a) μ = 45aph.

b) μ = 55aph.

c) μ = 70aph.

d) μ = 55aph. and 25 aph.

FIGURE 6-4 TERMINAL A. CUMULATIVE NUMBER OF USERS TURNED AWAY I(t); ARRIVAL AND HANDLING RATES AS IN FIGURE 6-1

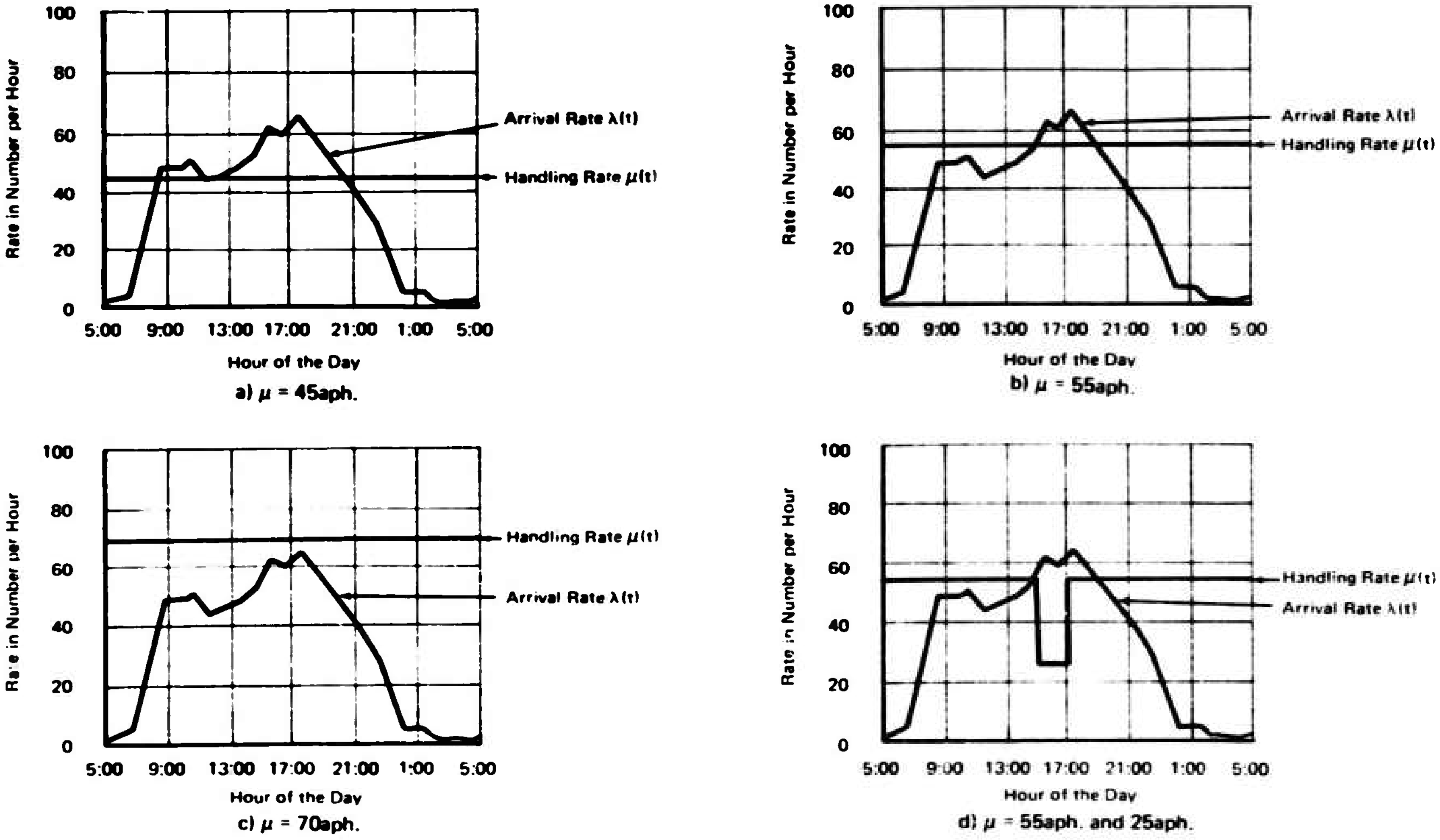


a) μ = 45aph.

b) μ = 55aph.

c) μ 70aph.

FIGURE 6-5 TERMINAL A. EXPECTED WAITING TIME W(t) IN MINUTES; ARRIVAL AND HANDLING RATES AS IN FIGURE 6-1a, b, c.



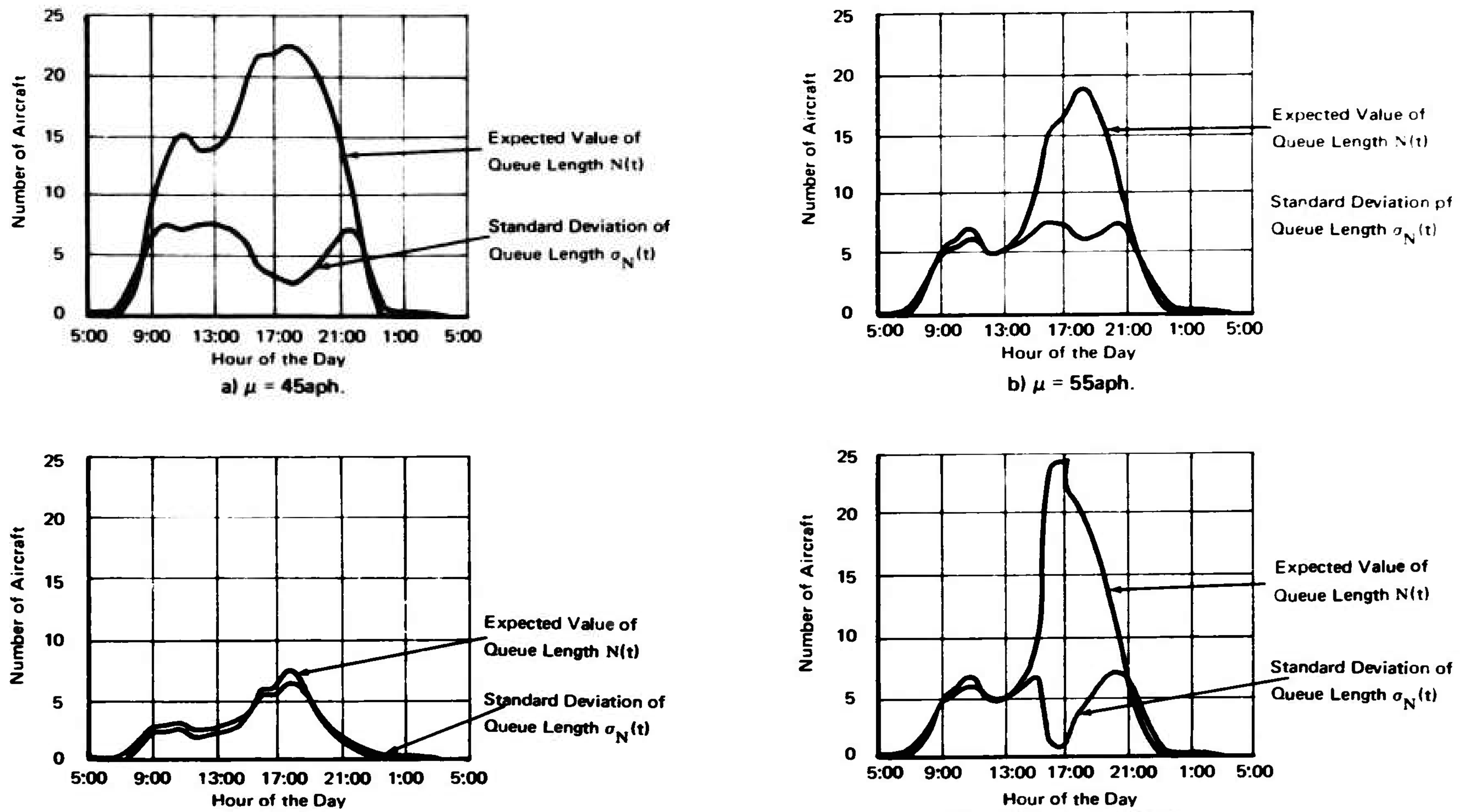
a) μ = 45aph.

b) μ = 55aph.

c) μ = 70aph.

d) μ = 55aph. and 25aph.

FIGURE 6-6 TERMINAL B. A TYPICAL OBSERVED ARRIVAL RATE $\lambda(t)$ AND FOUR ASSUMED HANDLING RATES $\mu(t)$



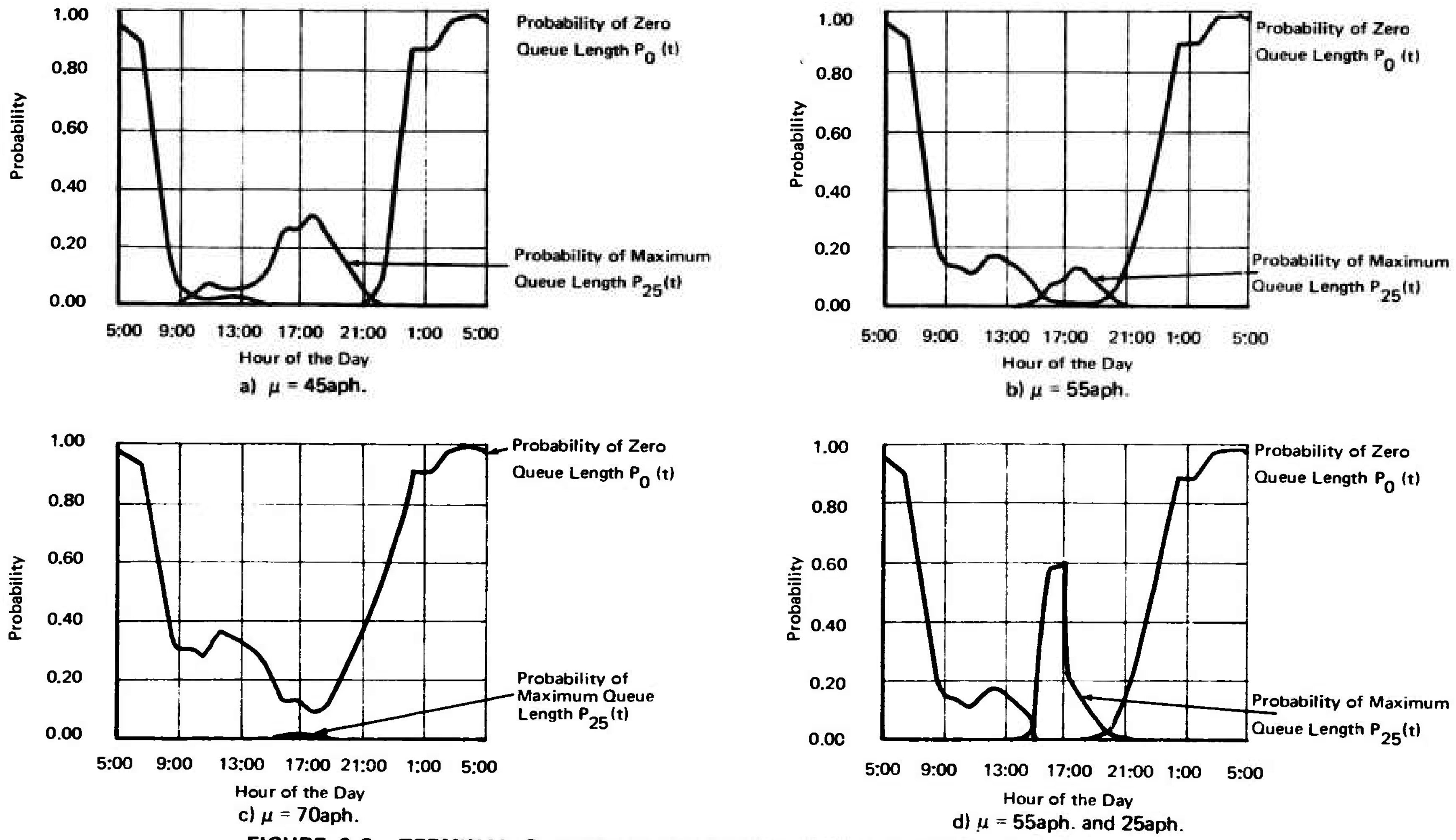
a) μ = 45aph.

b) μ = 55aph.

c) μ = 70aph.

d) μ = 55aph. and 25aph.

FIGURE 6-7 TERMINAL B. EXPECTED VALUE AND STANDARD DEVIATION OF QUEUE LENGTH; ARRIVAL AND HANDLING RATES AS IN FIGURE 6-6



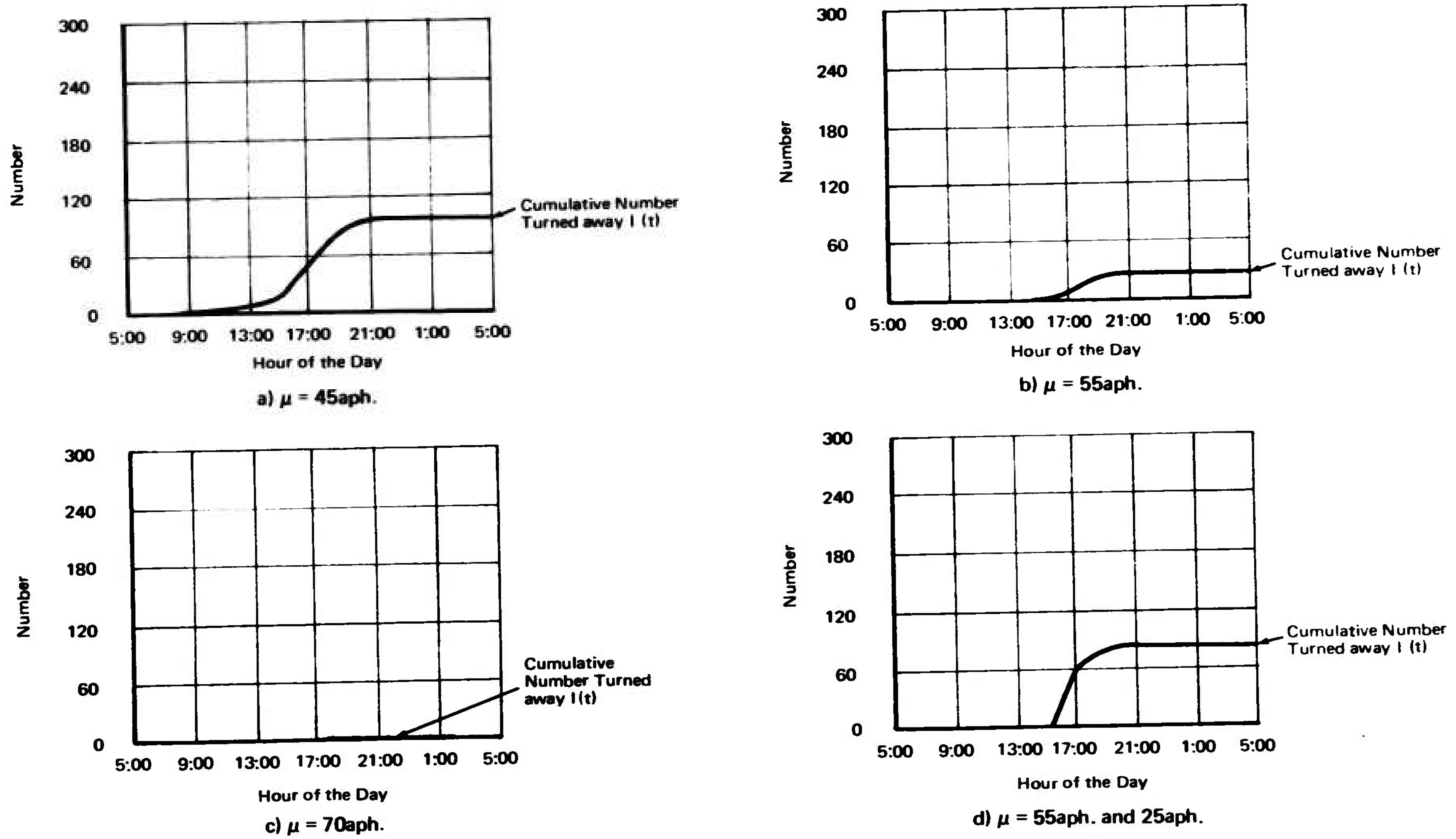
a) μ = 45aph.

b) μ = 55aph.

c) μ = 70aph.

d) μ = 55aph. and 25aph.

FIGURE 6-8 TERMINAL B. PROBABILITY OF MAXIMUM QUEUE LENGTH P_{25}(t) AND OF ZERO QUEUE LENGTH P_0(t); ARRIVAL AND HANDLING RATES AS IN FIGURE 6-6



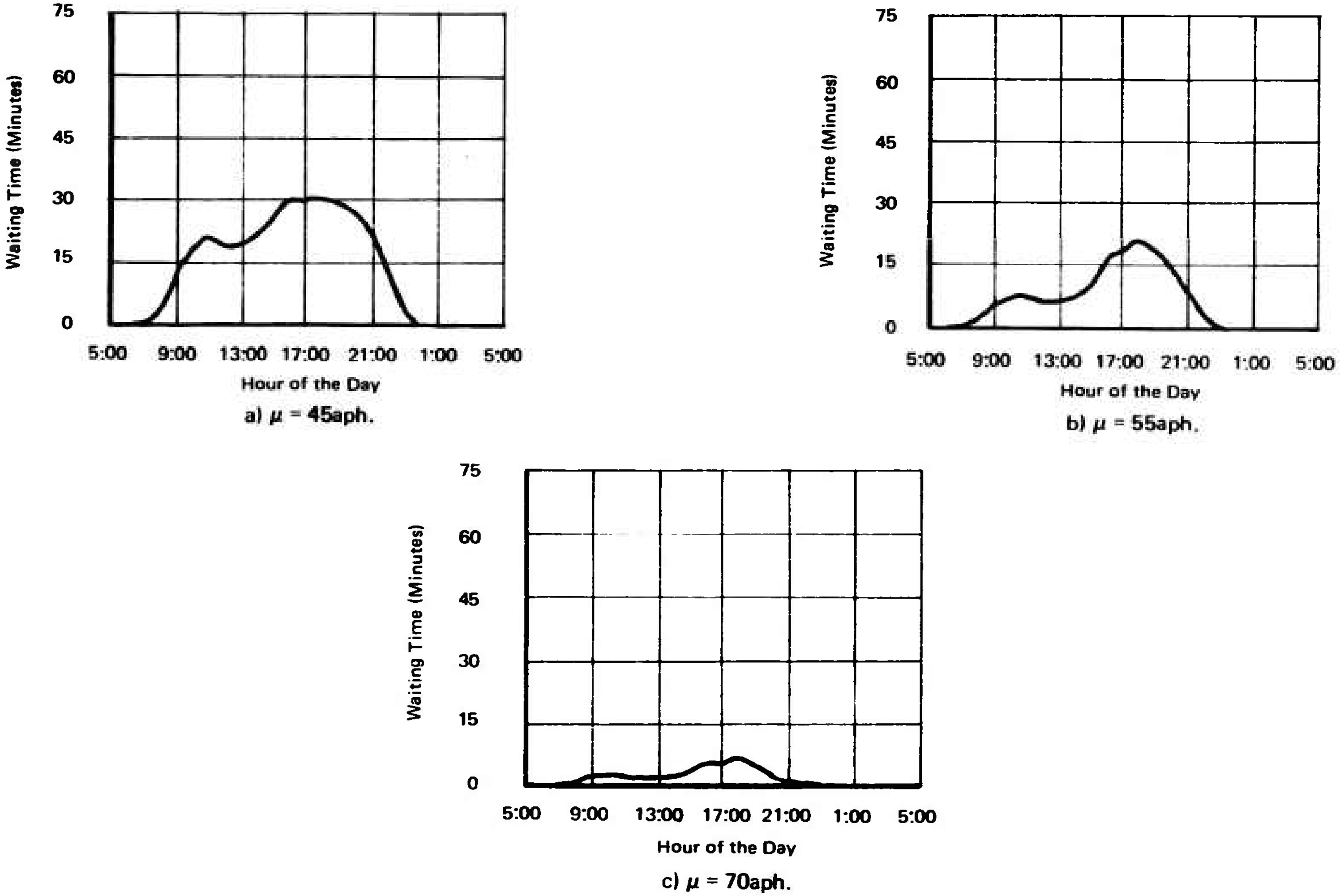
a) μ = 45aph.

b) μ = 55aph.

c) μ = 70aph.

d) μ = 55aph. and 25aph.

FIGURE 6-9 TERMINAL B. CUMULATIVE NUMBER OF USERS TURNED AWAY I(t); ARRIVAL AND HANDLING RATES AS IN FIGURE 6-6



a) μ = 45aph.

b) μ = 55aph.

c) μ = 70aph.

FIGURE 6-10 TERMINAL B. EXPECTED WAITING TIME W(t) IN MINUTES; ARRIVAL AND HANDLING RATES AS IN FIGURE 6-6a, b, and c.

which were obtained by the initial conditions $P_0(0) = 1, P_1(0) = \ldots P_m(0) = 0$, are plotted on the same scale, it is observed that they do not return to their original values after 24 hours: the functions $P_n(t)$ and the derived expected values, standard deviations, etc., are not periodic. In the case of Terminal B, on the other hand, they return to virtually their initial values at the end of 24 hours, so they will show up on the combined sequence of graphs as periodic functions, representing a *periodic steady state.*

This illustrates the theorem, proven mathematically in Section 6.4 (the same proof applies in the cases of Section 6.5), namely, that if the coefficients (inputs) are periodic, there *exists* one periodic solution; but this solution is unique and special, and the general behavior is not to be periodic, but to approach the periodic one with an increase of time.

Figures 6-1 through 6-10 present the results of the solutions to the four cases for each of the terminals, A and B. Figures 6-1 and 6-6 show the input data; the demand and service rates for the model (i.e., the arrival and handling rates) in numbers of aircraft per hour (aph). The results are plotted in all cases for the 24 hours between 5 a.m. and 5 a.m. In each of the four cases the arrival profile is identical; one profile for A and one for B. The handling rates in the four cases are 45, 55, and 70 aph (aircraft per hour), for the constant handling rate cases, and 55 aph with a drop to 25 aph during the period 3-5 p.m.

Figures 6-2 and 6-7 show the expected number of aircraft in the queue and the standard deviation of this number. Figures 6-3 and 6-8 show the probabilities of there being a full and an empty queue through the day. The curve with a maximum at the start and end of the 24-hour period is the curve of zero in the queue. Figures 6-4 and 6-9 show the cumulative number of aircraft turned away during the day. Figures 6-5 and 6-10 show the expected waiting time in the queue for those aircraft that are admitted to the queue, for the first three cases, only.

The input data, based on statistics taken at JFK and LaGuardia airports, represent the total operations per hour occurring at these airports. The model being used represents a single queuing system with a single service facility (which in reality may represent more than one runway) which can accept both "arrivals" and "departures" at a given constant rate. When we use the term "arrivals" in describing Figures 6-1 and 6-6, it imples that, given the origin of the data being used, a joint stream of "arrivals" and "departures" arriving at the service facility. Thus, the queue in this case does not represent a single physical entity, but rather the combination of the two queues: (a) in the air (located in holding areas), and (b) on the ground (located on a taxiway and held at the gates). In all the cases shown the queue is limited to 25 air and ground spaces in total.

These results clearly show the inappropriateness of a schedule in which up to 80 aircraft per hour may be expected to compete for only 25 aircraft spaces. This is clearly the cause of the very large number of aircraft that are turned away in one day (Figures 6-4 and 6-9). The assumptions give a perfectly good example of the operation of our analysis tools; in precise terms, these give a pessimistic answer to an unrealistically pessimistic question. In passing, it may be noted that in some cases a limitation in *endurance* rather than in number of queue *places* could be introduced into the formulation.

Secondly, the term "aircraft turned away," which represents the way that the model is set up, implies, to an excessive degree, the physical diversion of aircraft when they arrive at the holding fix or the end of the taxiway. This is unlikely to be a frequent happening, though the number turned away will reflect the number of flights that are cancelled or are diverted before they reach the terminal. Two side effects of this condition are excluded from the model: first, the flights cancelled will largely join the demand later during the 24-hour period if the overload is only temporary; and, second, the demand component represented by departures is clearly a function of the number of arrivals that can be accepted.

The graphs, to a large extent, speak for themselves. We can, however, note a few points of interest.

The maximum expected queue length occurs when the demand peak ends or when a service rate reduction is ended. It does not occur, as one might naïvely expect, when the demand profile is at its maximum value. This is entirely in accord with the observations cited in Section 3.2 of Chapter 3 on the capacity of an air terminal.

The standard deviation of the expected queue length (and, therefore, of the expected waiting time) is small when the queue is short, and also when the average queue length is near its maximum value. The standard deviation is greatest when the expected queue length is intermediate, and is even greater when the rate of change of the mean queue length is high. What does this mean? If both the demand and the service rate were constant, and if the maximum allowable queue length were infinite (or at least very large), we would expect an exponential distribution of probabilities as a function of queue length, in which case the standard deviation of the queue length would equal its mean. Thus, we can surmise that when the mean queue length and the rate of change of the mean queue length are both small, the standard deviation should approximate the mean. This is perhaps best illustrated in Figure 6-7c, or in Figure 6-2c, over the time period 0500 to 1400. By a similar argument, if a finite queue is nearly saturated, the standard deviation of the queue length should approximate the mean number of vacancies left in the queue. This is illustrated in Figure 6-2d, between 1500 and 2100, or in Figure 6-7b between 1500 and 1900.

In the rather unlikely event that the average queue length is about half its maximum allowable length, and is stable at that level for a considerable period of time, then the probabilities of the respective queue lengths must be nearly equal, and the standard deviation of the queue length should be approximately one-quarter of its maximum level. This can only occur if the demand and the handling rates are approximately equal for a considerable period of time. We can find such an interval between 1000 and 1400 in Figure 6-4d. In Figure 6-2d, we can see that the average queue length oscillates around a value of 12 or 13, and that the standard deviation is very stable at around 7.

It is also worth noting that a very rapid change in the mean value of queue length appears to be accompanied by a peak in the standard deviation. We see this, for example, in Figure 6-2c at 1600 and at 2200, and in Figure 6-7b at 1500 and 2000.

Generalization from these examples suggests that the uncertainty in queue length, as measured by standard deviation from the mean, varies with the queue length when the queue is very short, varies inversely with the queue length when the queue is nearly saturated, and has peaks when the population of the queue is in a state of rapid transition.

Comparing the handling rate patterns in Figure 6-4b and 6-4d, we can see that they differ only by the reduction of the handling rate from 55 to 25 aircraft per hour (aph) for a 2-hour period. Potentially, this results in a maximum reduction of 60 units of service. Comparing Figures 6-4b and 6-4d, we can see that the number turned away in 6-4d is almost exactly 60 greater than in 6-4b which is perfectly reasonable. However, it is surprising to note in Figures 6-9b and 6-9d that there is also a differential of almost exactly 60, in spite of the fact that the total number to be serviced in Figure 6-6d is quite modest in comparison with the total handling capacity. The story is told in Figure 6-7d. At time 1500, the average queue length is approximately 10, but within a very small fraction of an hour it rises to around 24, and from there until the end of the busy period at 1700, some 60 aircraft are turned away very rapidly. This simply shows that a queue which can store only 25 is incapable of providing reasonable carry-over when the excess of demand over service is 35 per hour for 2 hours.

We can compare the expected waiting time (Figures 6-5 and 6-10) with the daily mean waiting time shown in Table 6-1. This has been computed from the graphs and represents the mean height of the curve. We note that the mean waiting time throughout the day and the expected waiting time for the most active part of the day differ quite widely, the former being considerably less. The waiting time that could occur with some quite low probability, say $p = 0.1$, will be longer than the expected waiting time. This observation allows some light to be shed on the utility of the assumption of a mean delay (say of 4 minutes) as a measure of capacity. As can be seen, given quite realistic demand rates, such a blanket average can hide some quite significantly longer delays.

TABLE 6-1

DAILY MEAN WAITING TIME (minutes)

	Terminal A	Terminal B
Case 1	18.1	12.5
Case 2	10.9	6.2
Case 3	5.6	1.2

We exhibit these results not because of the specific numerical conclusions they suggest, but because they show transient phenomena which we believe are important and which cannot come out of steady-state analysis. We can see the delay of the peak mean queue length after the peak demand, the peaking of the uncertainty in queue length when the average queue length is in sudden transition, the extreme non-uniformity of waiting time during the day, and the failure of a short queue to hold over a peak demand for later service. This simple set of examples illustrates the flexibility which can be achieved and the potential which we believe this method has for experimentation and manipulation of queuing situations representative of ATC problems.

6.4 DIURNAL CYCLES IN LANDING QUEUES

6.4.1 The Problem. When a large and busy airport is faced, during hours of high traffic, with more arriving planes than it can handle – particularly during low visibility – planes are ordered into holding patterns to await their "servicing" turn; i.e., their permission to enter the landing pattern and, finally, to land. Thus a queue develops during such times. In extreme cases, as at Kennedy, the holding space itself may become filled to such a point that further planes are either held on the ground at their points of departure or are diverted to other landing fields. This situation represents an obvious case of degradation of utility: *no degradation* at times when there is no holding requirement; *delays* for planes that are obliged to hold; and more serious delays and *diversions* after the holding space is saturated. This section supplies a mathematical methodology for the quantitative study of these degradations under fairly general assumptions – more general than in most of the conventional methods.

6.4.2 The General Mathematical Situation. One of the most obvious facts in the situation described above is that it is strongly time-dependent: during the "rush hour" periods (8-10 a.m. and 4-6 p.m.) there is a high volume of traffic, with the possible development of considerable holdings and delays. At the other extreme, during the night and early morning hours, there is very little demand for landing service, and aircraft arriving at the airport will be virtually certain of being allowed to land without holding. Finally, the pattern of arrivals is, to a good approximation during weekdays, *periodic* with period T = 24 hours.

Therefore we are faced with a queuing line problem, but one in which the conditions are strongly time-dependent and periodic. A *steady-state* solution would, therefore, involve a contradiction, so that the usual treatments are completely ruled out. On the other hand, a *periodic* solution (the rising and falling of queues with the passage of time and within a 24-hour period) takes its place, and can give important quantitative indications of the state of affairs.

It follows that the conventional treatment of the queues in the airport landing problem, in which a stationary solution of a time-independent waiting line situation is obtained, is altogether beside the point – whethey they are carried out analytically or by a computer simulation.

In the present study of the problem, a one-step Markov process will be assumed in which the transition probabilities are strongly varying periodic functions of the time with period T = 24 hours. It will be shown that, *whatever these transitional probabilities may be,* there exists a unique periodic solution of period T. This, of course, plays the role of the steady-state solution of the time-independent transition processes of conventional operational studies.

In only one treatment known to the author is the assumption of time-independent transition probabilities avoided.[2] The method used in that paper is computer simulation, which is incapable, *in principle,* of disclosing such general facts as the existence and uniqueness of a periodic solution of the stochastic equations. Indeed, no use is made by these authors of the simple mathematical methods that have been available for decades for handling their problems.

6.4.3 Specific Assumptions. To obtain a mathematical treatment, at once realistic and tractable, we start with the basic assumption that the time necessary for an aircraft which enters the landing pattern to land is *short* (e.g., 1 minute) in comparison with the time interval (e.g., 30 minutes) during which the arrival rates change appreciably. Consequently, an intermediate interval of time h exists, for which the following statements are approximately true:

During the time interval (t, t + h) the probability that an aircraft, next in line for landing, be permitted to enter the landing pattern is Lh = L(t)h. Actually, we shall assume L(t) independent of t, so that it equals the μ of conventional "birth-death" processes.

During the same time interval (t, t + h), the probability that an aircraft adds itself to the holding (or landing) pattern is R(t)h, and R(t + T) = R(t). This R would, if independent of t, be the transition probability λ of the "birth-death" or the Poisson process. This assumes that there is room in the holding pattern; when there is not, the probability of the addition is *zero.*

The state of the system at any given time is completely characterized by the number N_t of aircraft in the total system: the *holding plus landing patterns.*

The total capacity of these is the fixed number m: $N_t = 0, 1, \ldots, m$.

To quantities of order h, the only probabilities of transitions of N_t are to its immediate neighbors: *two* neighbors from a value $N_t = n$ when $0 < n < m$; *one* when $n = 0$ or $n = m$.

On introducing

$$P_n(t) = \text{prob}\left\{N_t = n\right\}$$

we have, to quantities of order h,

$$P_n(t \;+ h) = P_{n-1}(t)\,Rh + P_n(t)[1 - Rh - Lh] + P_{n+1}(t)\,Lh,$$

provided $0 < n < m$. Obvious modifications are made when $n = 0$ or $n = m$. To maintain a greater flexibility, we shall carry the general treatment through in the case in which the transition probability coefficients R and L depend not only on t but on the state n out of which they occur. Then the above equation becomes:

$$P_n(t + h) = P_{n-1}\; R_{n-1}h + P_n[1 - R_n h - L_n h] + P_{n+1}\; L_{n+1}h;$$

and, similarly, for $n = 0$, $n = m$. It is understood that the capital letters on the right are functions of t.

We now make the approximation, which expresses our first assumption, that

$$P_n(t + h) = P_n(t) + P'_n\,(t)\,h,$$

the error being of the order of h^2. The above equation, on dropping such higher order terms, becomes the middle one in the following system of m + 1 equations in m + 1 unknown functions:

$$P'_0(t) = -\,R_0(t)\,P_0(t) + L_1(t)\,P_1(t)$$

$$P'_n(t) = R_{n-1}(t)\,P_{n-1}(t) - [R_n(t) + L_n(t)]P_n(t) + L_{n+1}(t)\,P_{n+1}(t)\;(0 < n < m) \tag{6-1}$$

$$P'_m(t) = R_{m-1}(t)P_{m-1}(t) - L_m(t)P_m(t)$$

$$R_i(t+T) = R_i(t)$$

$$L_i(t+T) = L_i(t) \qquad (6\text{-}2)$$

6.4.4 Application of the General Theory. Equations (6-1) form a homogeneous linear differential system of order $m + 1$: that is, $m + 1$ first-order equations in the $m + 1$ functions $P_n(t)$ ($n = 0, 1, \ldots, m$). By the general theory of ordinary differential equations (the coefficients being assumed continuous for all $t \geq 0$), there exists one and only one solution taking on the pre-assigned initial values $P_n(0) = c_n$. Furthermore, on adding all $m + 1$ Equations (6-1), we find the value *zero* on the right, while on the left, $P_0'(t) + P_1'(t) + \ldots + P_m'(t)$. Therefore, we have:

$$\frac{d}{dt}\left[P_0(t) + P_1(t) + \ldots + P_m(t)\right] = 0,$$

so that $P_0(t) + P_1(t) + \ldots P_m(t)$ is a constant equal to its initial value $c_0 + c_1 + \ldots + c_m$. Assuming that the latter is unity, we have that for all $t \geq 0$:

$$P_0(t) + P_1(t) + \ldots + P_m(t) = 1$$

which is one of the basic properties of a probability distribution. The second property $P_i(t) \geqq 0$ is a less immediate consequence of the general theory. It is easily established by going back to one of the basic algorithms of the latter: the method of *successive approximations*, applied in a particular form.

If in Equations (6-1) $P_{n-1}(t)$ and $P_{n+1}(t)$ are regarded as known, each equation becomes one of the first order, which can be solved by quadratures, after transposing the term in $P_n(t)$ and multiplying through by the integrating factor

$$\exp \int_0^t \left[R_n(t) + L_n(t)\right] dt$$

$$(0 < n < m)$$

(When $n = 0$, $L_0(t)$ is dropped; when $n = m$, $R_m(t)$ is dropped.) This makes the left-hand member an exact derivative. On integrating this and making an obvious division, and so forth. Equations (6-1) yields the following system of (Volterra) integral equations, in which we use the abbreviation:

$$G_n(t) = R_n(t) + L_n(t) \qquad (0 < n < m) \qquad (6\text{-}3)$$

$$= R_0(t) \qquad (n = 0)$$

$$= L_m(t) \qquad (n = m)$$

$$P_n(t) = c_n \exp\left[-\int_0^t G_n(t)\,dt\right] \qquad (6\text{-}4)$$

$$+ \int_0^t \exp\left[\int_t^s G_n(t')\,dt'\right] \cdot \left[R_{n-1}(s)\,P_{n-1}(s) + L_{n+1}(s)\;P_{n+1}(s)\right] ds$$

for $0 < n < m$, and where $L_0(s)$ or $R_m(s)$ are dropped on the right when $n = 0$ or m.

These equations, which are completely equivalent to (6-1), are solved by successive approximations, using the following schema:

$$P_n^0\,(t) = c_n \exp\left[-\int_0^t G_n(t)\,dt\right]$$

$$P_n^{k+1}(t) = c_n \exp\left[-\int_0^t G_n(t)\,dt\right] \qquad (6\text{-}5)$$

$$+ \int_0^t \exp\left[\int_t^s G_n(t')\,dt'\right]$$

$$\bullet\left[R_{n-1}\,(s)\,P_{n-1}^k(s) + L_{n+1}\,(s)\,P_{n+1}^k(s)\right]\;ds$$

$$k = 0, 1, \ldots j \qquad (0 < n < m)$$

with the appropriate modifications for $n = 0$ and m.

Since the initial probabilities c_i are non-negative, (6-5) shows by induction that no $P_n^k(t)$ can be negative. As a matter of fact, since we are assuming that *some* aircraft enter the system and land in each 24-hour period, while for some values of t, $G_n(t)$ may vanish, we must have:

$$\int_0^T G_n(t)\,dt > 0.$$

Hence every $P_n^k\ (T) > 0$.

Standard elementary methods show the convergence of the sequence $P_n^k\ (t) \rightarrow P_n\ (t)$ as $k \rightarrow \infty$ and that this limit satisfies (6-4), and hence (6-1) and the initial conditions. Therefore, we have shown that the unique solution of the equations in question is a probability distribution $P_n\ (t)$, and that $P_n\ (T) > 0$.

We now turn to the question of periodicity. Let (6-1) be written with t replaced by t + T. Since each coefficient, $R_n\ (t)$ and $L_n(t)$ has T as a period

$$R_n(t + T) = R_n(t), \qquad\qquad L_n(t + T) = L_n(t),$$

it is seen that *the system* $\{P_n(t + T)\}$ $(n = 0, 1, \ldots, m)$ *satisfies* (6-1).

Let $P_{kn}\ (t)$ be the set of solutions of (6-1) determined by the initial values:

$$P_{kn}(0) = 0 \qquad \text{when } k \neq n, \qquad P_{nn}(0) = 1. \tag{6-6}$$

Being linearly independent, these form a *fundamental system* of m + 1 solutions of (6-1), in terms of which any other solution can be expressed as a homogeneous linear combination with constant coefficients. It is, in fact, evident that our earlier solution $P_n(t)$ with the given initial values c_n is given by

$$P_n(t) = \sum_{k=0}^{m} c_k\ P_{kn}(t).$$

Now replace t by t + T. The new solution $P_n\ (t + T)$ must also be a linear combination of $P_{kn}\ (t)$:

$$P_n(t + T) = \sum_{k=0}^{m} c'_k\ P_{kn}(t).$$

As a matter of fact, since we have

$$P_n(t + T) = \sum_{k=0}^{m} c_k\ P_{kn}(t + T),$$

and since $P_{kn}\ (t + T)$ represents a set of m + 1 solutions,

$$P_{kn}(t+T) = \sum_{j=0}^{m} a_{kj}\, P_{jn}(t), \tag{6-7}$$

we have

$$P_n(t+T) = \sum_{k=0}^{m} c_k \sum_{j=0}^{m} a_{kj}\, P_{jn}\,(t).$$

Comparing this with the earlier expression for $P_n(t + T)$ and invoking the fact that the coefficients in the linear expression of a solution in terms of a fundamental system are unique, we find:

$$c_j' = \sum_{k=0}^{m} c_k\, a_{kj} \tag{6-8}$$

The elements a_{kj} of this matrix are obtained at once from (6-7) on setting t = 0:

$$P_{kn}\,(T) = a_{kj} \tag{6-9}$$

From what we have shown above, the matrix a_{kn} is a *stochastic matrix* of the simplest sort, i.e.:

$$a_{kn} > 0; \qquad \sum_{n=0}^{m} a_{kn} = 1.$$

We shall denote it in matrix notation by $A = (a_{kn})$. It is a matrix of transition probabilities from states at t = 0 to those at t = T. Because of the T-periodic nature of the basic differential equations, it is also the transition probability matrix from the states at any epoch t to the congruent epoch t + T.

As a last application of the general theory,[3] we know that for such a transition matrix, there exists one and only one invariant vector C: $(c_0, c_1, \ldots, c_m)$ of positive numbers adding up to unity ($c_i > 0$; $\Sigma c_i = 1$), invariant in the sense that, for it (6-8) gives $c_j' = c_j$; i.e., it is a left eigenvector with unit eigenvalue:

$$\sum_{k=0}^{m} c_k\, a_{kj} = c_j$$

or, equivalently, C A = C.

In view of the earlier equations, this is the necessary and sufficient condition that the probability distribution $P_n(t)$ having these c's as initial values (or values at $t = t_0$) be periodic with the period T. Thus we have proved:

> **There is a unique periodic solution of the system of stochastic equations (6-1).**

A further fact is derived from the ergodic properties of the transition matrix A:

> **Every probability distribution satisfying (6-1) approaches the above periodic solution as t increases indefinitely (and does so at a geometric rate).**

The standard theory[4] shows that this is true for the epochs t = 0, T, 2T, ... Since the values in each interval $sT \leqq t \leqq (s + 1) T$ are determined as continuous functions of their values at the extremity t = sT, it follows that these intermediate values approach those of the periodic solution (and uniformly) as the interval moves out indefinitely.

In view of these facts, the periodic solution can be regarded as the state of kinetic equilibrium or periodic steady state of the system, playing the role of the constant steady-state solution in the case of time-independent transitions.

In the next section its relationship with delay and holding times will be investigated.

6.4.5 Waiting Times. Let us suppose that a particular aircraft enters the system (the holding + landing pattern) at the epoch t and becomes the n'th member of the waiting line (n = 1, 2, ..., m); and assume further that there is a fair rule of service: first in line-first landing, etc. This aircraft will land at a later epoch t', where $\tau = t' - t$ is the length of time up to the moment when a total of n landing opportunities occur.

In the usual case it is sufficiently accurate to assume that all the leftward transition probability coefficients are equal and independent of the time.

Setting

$$L_1 = L_2 = \ldots = L_m = \mu$$

we observe that a *gain* in place of s units, i.e., the reduction of n to n - s (s $\leqq$ n) is a Poisson process. Therefore the probability that s has the value n − 1 after the time τ is

$$\frac{(\mu\tau)^{n-1}}{(n-1)\,!}\; e^{-\mu\tau} .$$

Therefore the probability of s reaching the value n precisely during the interval of wait $(\tau, \tau + d\tau)$ is, to quantities of the first order, the above expression multiplied by $\mu d\tau$. The expected wait W_n, given the arrival at the n'th place, is:

$$W_n = \int_0^\infty \frac{(\mu\tau)^{n-1}}{(n-1)\,!}\; e^{-\mu\tau}\; \mu\tau\; d\tau = \frac{n}{\mu} . \qquad (6\text{-}10)$$

If the landing pattern can hold ℓ aircraft, and each is an average of 2 minutes apart, then the time taken in the landing pattern by each plane is 2ℓ minutes, or, equating this to the above expected value, $\mu = 1/2$. The period of 2ℓ minutes being regarded as the minimal (i.e., nondelay) time, anything further can be regarded as a *delay* due to holding. From formula (6-10), its expected value is $W_n - 2\ell$ or, generally,

$$W_n - \frac{\ell}{\mu} = \frac{n-\ell}{\mu} . \qquad (6\text{-}11)$$

This simple expression assumes known the place of our aircraft in the holding queue when it arrives there. To obtain more generally applicable information, showing the effect of varying numbers in the holding pattern, we take the expected value of W_n over n; i.e., we compute:

$$W(t) = \sum_{n=1}^{m} W_n\, P_n(t) = \frac{1}{\mu} \sum_{n=0}^{m} nP_n(t) = \frac{1}{\mu}\, \bar{N}_t \qquad (6\text{-}12)$$

The computation is simplified when we return to the original assumption concerning the rightward transition probability coefficients:

$$R_0 = R_1 = \ldots = R_{m-1} = \lambda(t)$$

where $\lambda(t)$ is a given periodic function of t and represents the coefficient of Poisson arrival of aircraft.

From Equations (6-1) we obtain, by multiplying the n'th by n and adding from n = 0 to n = m,

$$\frac{d}{dt}\ \bar{N}_t = \lambda(t)\,[1 - P_m(t)] - \mu\,[1 - P_0(t)] \qquad (6\text{-}13)$$

This is integrated and inserted in (6-12). The result can be obtained numerically once $\lambda(t)$ and μ are known, the Equations (6-1) solved numerically to get $P_{ij}(t)$ and hence a_{ij}, then the eigenvector giving the values of $(c_0, c_1, \ldots, c_m)$ for the periodic solution found by determinant calculations, and finally, the numerical values of $P_m(t)$ and $P_0(t)$ calculated. The formal result at this point is

$$W(t) = W(0) + \int_0^t \left\{ \frac{\lambda(t)}{\mu}\ [1 - P_m(t)] - [1 - P_0(t)] \right\} dt. \qquad (6\text{-}14)$$

If the situation is such that there is always a time (e.g., in the middle of the night) when there are no planes in the system, and if that time is taken as t = 0, we shall have $W(0) = 0$ in (6-14). In every case, however, $W(t)$ is periodic with period T.

Assuming, as we may, that the functions involved [$\lambda(t)$, and therefore all the $P_n(t)$] are continuously differentiable, the maximum waiting time will occur when $W'(t) = 0$, $W''(t) < 0$. Using (6-13) and the result of differentiating through and then applying (6-1), we obtain explicit conditions for a maximum.

Similar but more complicated methods yield the standard deviation of $W(t)$. However it is more convenient for this purpose to introduce the probability generating function.

6.4.6 The Generating Function. This is the m'th degree polynomial in x, with coefficients functions of t, defined by the equation:

$$g = g(t,x) = P_0(t) + P_1(t)x + \ldots + P_m(t)x^m. \qquad (6\text{-}15)$$

In the case of constant left and right translation coefficients, a differential equation for it is easily derived from (6-1), by multiplying the n'th equation by x^n and adding. We obtain:

$$\frac{\partial g}{\partial t} = (\lambda x + \frac{\mu}{x} - \lambda - \mu)g + \mu(1 - \frac{1}{x})P_0 + \lambda x^m (1 - x)P_m \tag{6-16}$$

This is linear of the first order in t, and could be solved by quadratures if P and P_m were known; then every P_n would be determined as explicit expressions.

The present use of (6-16) is to obtain the moments; for we have for the k'th moment formula:

$$\overline{N_t^k} = \left[\left(x \frac{\partial}{\partial x}\right)^k g\right]_{x=1}$$

When k = 0, this gives g(t, 1) = 1, while for k = 1 and 2 we obtain

$$\frac{d}{dt}\overline{N}_t = \lambda\,[1 - P_0(t)] - \mu\,[1 - P_m(t)] \tag{6-17}$$

as before; and

$$\frac{d}{dt}\overline{N_t^2} = \lambda + \mu + 2(\lambda - \mu)\overline{N}_t - \mu P_0 - \lambda(2m + 1)P_m \tag{6-18}$$

For the standard deviation, $\sigma^2 = \overline{N_t^2} - (\overline{N}_t)^2$, we see that

$$\frac{d}{dt}\sigma^2 = \frac{d}{dt}\overline{N_1^2} - 2\overline{N}_t\frac{d\overline{N}_t}{dt} \tag{6-19}$$

$$= \lambda + \mu - \mu P_0\,[1 + 2\overline{N}_t] - \lambda P_m\,[2m + 1 - 2\overline{N}_t];$$

from which σ^2 can be obtained by the integration of known functions [$\overline{N}_t$ having previously been obtained from (6-17)].

6.5 A MATHEMATICAL FORMULATION OF THE INTERACTION OF LANDING AND TAKE-OFF QUEUES UNDER TIME-DEPENDENT CONDITIONS

6.5.1 The Problem. We consider an airport with a single runway. This runway is to be used by aircraft landing at the airport and those taking off. During hours of high traffic, with more than one plane desiring the use of the runway, planes are ordered into holding patterns or queues to await their turn for landing. Similarly, a queue is formed on tne ground consisting of aircraft awaiting their turn for take-off. In queuing terminology we have a system of two types of customers forming two separate queues for service by a single server. In this section we will refer to the two queues as arrival (or landing) and departure (or take-off) queues.

The characteristic property of such queues, which distinguishes them from those studied in so much of conventional queuing theory, is that the circumstances of their operation may be strongly *time-dependent.* This is because the rates at which the aircraft arrive at the landing queue, and also at the take-off queue, may be much higher at certain times of the day (rush hours) than at others (early morning). Therefore, the problem has to be solved for time-dependent input parameters (which appear as given functions of the time in the differential difference equations of the process). In many cases we may assume that these have a diurnal periodicity (a 24-hour period).

The waiting line problem is that all queues *are constrained not to exceed given lengths.* This constraint, which is intended to reflect the limited air and ground space near the terminal, has the effect of dispensing with the infinite (or indefinitely growing) queues of conventional treatments, and of leading to *finite* systems of linear differential equations.

We are interested in studying the waiting time for an aircraft through the system, the number of aircraft "lost" in a given time, the expected number of aircraft in the system at time t, and various other statistics of the queuing system, related to its capacity.

This section represents the first step toward the solution of the above queuing system: it formulates, under rather general assumptions, the differential difference equations determining the queue behavior, and associated with the various queue disciplines relevant to air traffic problems.

6.5.2 Assumptions. The rule governing the arrival of aircraft into each of the two queues is such that the aircraft arrive at "random," the number of arrivals in time t being a Poisson variable, and the time interval between two consecutive arrivals having the exponential density. The parameters of the Poisson process associated with the departure (take-off) and arrival (landing) queues are represented by λ and λ', respectively, both of which are functions of time. For example:

$\lambda(t)$ = rate at which aircraft join the departure Q (at time t). Similarly,
$\lambda'(t)$ = rate at which aircraft join the arrival Q (at time t).

Next, we assume that the service time distributions for aircraft in the two queues are exponential with parameters dependent on the number of aircraft in the two queues, but *independent of time.*

μ_{ij} = service rate for aircraft from departure queue when there are i aircraft in the arrival queue and j aircraft in the departure queue. Similarly, μ'_{ij} is the corresponding rate for the arrival queue.

6.5.3 Possible Queue Disciplines. When the maximum number of aircraft allowed in the landing and take-off queues is m and n respectively, we consider the following three types of priority disciplines:

- *Strict Arrival Priorities* – Here aircraft in the arrival queue have priority over aircraft waiting in the departure queue for use of the runway. The priority discipline is however, nonpreemptive; i.e., an aircraft from the take-off queue uses the runway when there are no aircraft in the landing queue, but it is allowed to complete its "service" in the event of an arrival into the landing queue.

- *Alternating Priorities* – In this case, aircraft from the two queues are serviced alternately, the selection for service within each queue being strictly on a first-come/first-served basis.

- *Mixed Priorities* – Here the priority rule is determined as follows: for j = number of aircraft in the take-off queue and for r an integer such, that $0 < r < n$, then (1) if $0 < j < r$, the arriving aircraft have priority over the departing aircraft; (2) if $r \leqslant j \leqslant n$, then departing aircraft have priority over arriving aircraft.

6.5.4 Transition Equations for the Various Queues. The state of the system at any time t is represented by the number i of aircraft in the arrival queue (i is positive, or zero when this queue is empty), the number j in the departure queue ($j \geqslant 0$), and also by whether the aircraft being served (using the runway) at t is a landing aircraft or an aircraft taking off; we shall label the former case with the index $k = 1$ and the latter with $k = 2$. Thus the state of the system is, for present purposes, fully described by the three indices (i, j, k) where $0 \leqq i \leqq m$, $0 \leqq j \leqq n$, $k = 1$ or 2. Finally, we shall denote by $P_{i\ j}(t)$ and $Q_{i\ j}(t)$ the probabilities that it be in the state $(i, j, \kappa = 1)$ and $(i, j, k = 2)$, respectively. Note that no state exists corresponding to $(i, 0, 2)$, $(0, j, 1)$; among $2(n + 1)(m + 1)$ different symbols (i, j, k) only

$$2(n + 1)(m + 1) - n - m - 2 = 2nm + n + m$$

correspond to *states* of the system. There is one additional state: when there is no aircraft in either waiting line, none being served, and k is undefined. We call this probability of this state $R(0,0)$. The range of allowable combinations of i, j, and k is shown in Figure 6-11. The symbols $P_{i,j}(t)$ and $Q_{i,j}(t)$ for the nondefined "states" are conventionally defined to be zero.

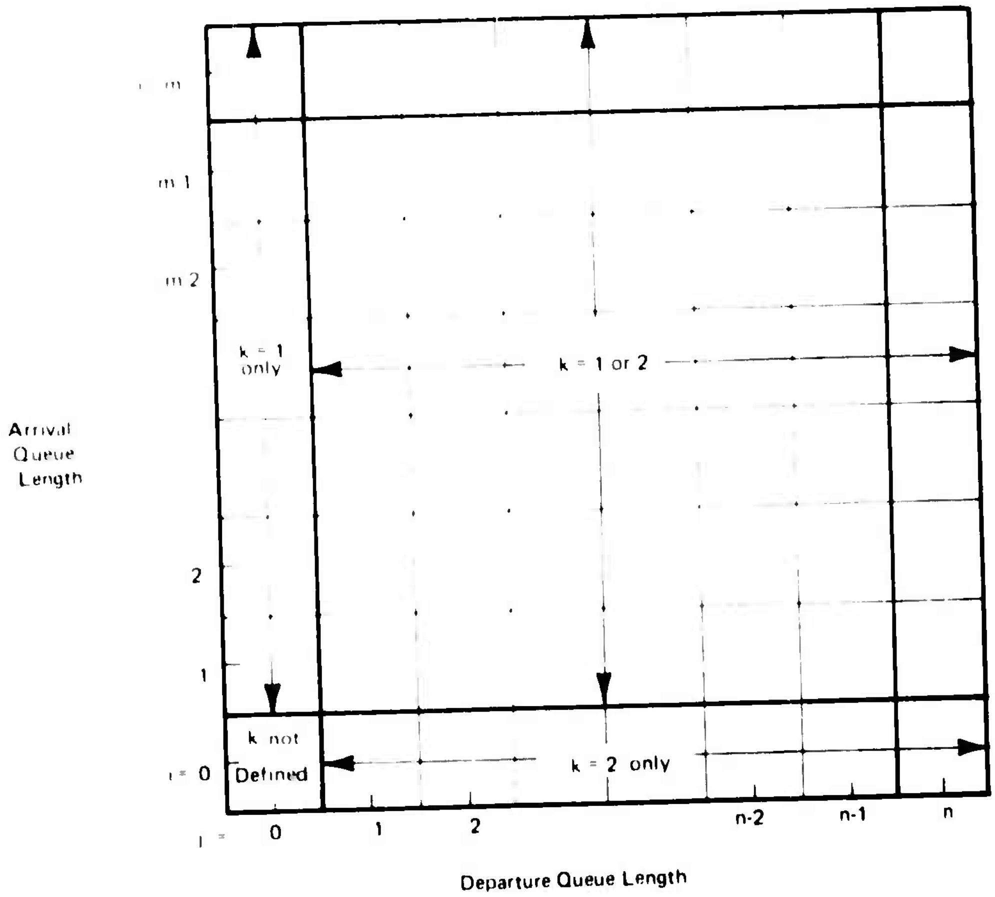


FIGURE 6-11 RANGE OF ALLOWABLE VALUES OF i,j,and k

Let us enumerate the ways in which we can arrive in a state where the arrival queue has length i and the departure queue has length j. The enumeration is facilitated by the diagram in Figure 6-12. If no plane completes a landing or takeoff, and no plane joins an arrival queue or departure queue, there is no change: this is the transition from queue lengths i, j to queue lengths i, j (this is a transition in the same sense that 0 is a number). In this case, k does not change. If an aircraft joins the arrival queue, the arrival queue length increases from i − 1 to i, and k does not change. If an aircraft joins the departure queue, the departure queue increases from j − 1 to j, and again k does not change. If an aircraft completes a landing, the arrival queue decreases from i + 1 to i, and the value of k must have been 1. If an aircraft completes a take-off, the departure queue length decreases from j + 1 to j, and the value of k must have been 2. In the latter two instances, the new value of k to be adopted depends on the priority rules.

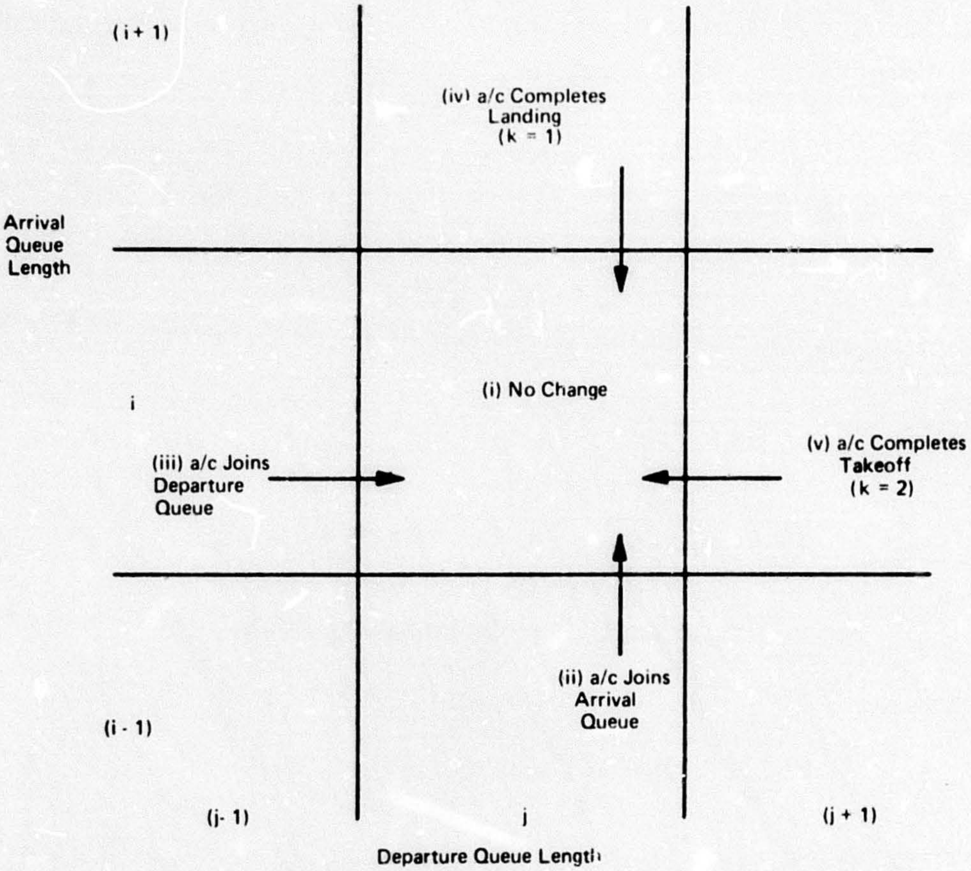

(i + 1)
(iv) a/c Completes Landing (k = 1)
Arrival Queue Length
(i) No Change
i
(iii) a/c Joins Departure Queue
(v) a/c Completes Takeoff (k = 2)
(ii) a/c Joins Arrival Queue
(i - 1)
(j- 1)
j
(j + 1)
Departure Queue Length

FIGURE 6-12 THE FIVE POSSIBLE TRANSITIONS LEADING TO ARRIVAL QUEUE LENGTH i, DEPARTURE QUEUE LENGTH j

For reasons considered below, we will ignore transitions requiring two simultaneous changes.

Now let us examine the effects that these transitions have upon the state probabilities. For the sake of definiteness, suppose that we examine the transitions to a state where the arrival queue length is i, $1 \leqq i \leqq m - 1$, and the departure queue has length j, $1 \leqq j \leqq n - 1$, and $k = 1$ (that is, an aircraft is landing). We use the same basic assumptions as in Section 6.4, that is, the service time for either landing or take-off is short in comparison with the interval of time required for an appreciable change in arrival or departure rate. Then, treating the various rates as nearly constant for a short time interval Δt, the probabilities of the five possible transitions are as follows:

(i) At time t the state of the system is (i, j, 1) with probability $P_{ij}(t)$ and during $(t, t + \Delta t)$ there is no arrival into either queue and the aircraft using the runway at time t does not complete its use of the runway. The probability of this event is:

$$P_{ij}(t)\,[1 - \lambda(t)\Delta t]\,[1 - \lambda'(t)\Delta t]\,[1 - \mu'_{ij}\Delta t].$$

(ii) At time t the system is in state (i − 1, j, 1) and during $(t, t + \Delta t)$ an aircraft joins the arrival queue. Everything else remains the same. The joint probability of this event is:

$$P_{i-1,j}(t)\,\lambda'(t)\Delta t\,[1 - \lambda(t)\Delta t]\,[1 - \mu'_{i-1,j}\Delta t].$$

(iii) At time t the state is (i, j − 1, 1) with probability $P_{i,j-1}(t)$ and during $(t, t + \Delta t)$ an aircraft joins the departure queue with probability $\lambda(t)\Delta t$ and there is no other change in the system. The corresponding joint probability is

$$P_{i,j-1}(t)\;\lambda(t)\Delta t\,[1 - \lambda'(t)\Delta t]\,[1 - \mu'_{i,j-1}\Delta t].$$

(iv) At time t the state is (i + 1, j, 1) and during $(t, t + \Delta t)$ the aircraft from the arrival queue using the runway completes its use of the runway with probability $\mu'_{i+1,j}\Delta t$ and there are no other changes in the system. The joint probability of this event is:

$$P_{i+1,j}(t)\,[1 - \lambda(t)\Delta t]\,[1 - \lambda'(t)\Delta t]\,\mu'_{i+1,j}\Delta t.$$

(v) At time t the system is in state (i, j + 1, 2) with probability $Q_{i,j+1}(t)$ and during (t, t + Δt) an aircraft from the departure queue completes its use of the runway with probability $\mu_{i,j+1}\Delta t$). The joint probability of this event is $Q_{i,j+1}(t)\,\mu_{i,j+1}\Delta t$.

(vi) Probability of more than one event happening in time Δt is $O(\Delta t^2)$. We shall not enumerate these probabilities.

Ignoring terms in which Δt appears in powers greater than 1, we have the following transient equation for our system for different values of i and j:

$$
\begin{aligned}
P_{ij}(t+\Delta t) &= P_{ij}(t)\,[1 - \lambda(t)\Delta t - \lambda'(t)\Delta t - \mu'_{ij}\Delta t] \\
&+ P_{i-1,j}(t)\,\lambda'(t)\Delta t + P_{i,j-1}(t)\lambda(t)\Delta t \\
&+ P_{i+1,j}(t)\mu'_{i+1,j} + Q_{i,j+1}(t)\mu_{i,j+1}\Delta t \quad \text{for } 0 < i < m \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad 0 < j < n
\end{aligned}
$$

On multiplying, transposing $P_{ij}(t)$, $(0 < i < m,\ 0 < j < n)$ to the left and dividing by Δt, the equation becomes:

$$
\frac{P_{ij}(t+\Delta t) - P_{ij}(t)}{\Delta t} = -\,[\lambda(t) + \lambda'(t) + \mu'_{ij}]\,P_{ij}(t
$$

$$
+ P_{i-1,j}(t)\,\lambda'(t) + P_{i,j-1}(t)\lambda(t) + P_{i+1,j}(t)\,\mu'_{i+1,j} + \mu_{j+1}\,Q_{i,j+1}(t)
$$

If we take limits as Δt→0, then, by definition, the left side is the derivative $dP_{ij}(t)/dt$, etc., and the equation becomes:

$$\frac{d}{dt} P_{ij}(t) = -[\lambda(t) + \lambda'(t) + \mu'_{ij}] P_{ij}(t) + \lambda'(t)P_{i-1,j}(t)$$

$$+ \lambda(t) P_{i,j-1}(t) + \mu'_{i+1,j} P_{i+1,j}(t) + \mu_{i,j+1} Q_{i,j+1}(t) \qquad \begin{matrix} 0 < i < m \\ 0 < j < n \end{matrix} \qquad (6\text{-}20)$$

This illustrates the procedure for deriving a differential equation from the assumptions about transition and probability. There is a similar differential equation for the complementary probability $Q_{i,j}(t)$. Furthermore, when i has one of its extreme values, 0 or m, or j has one of its extreme values, 0 or n, certain of the transitions are forbidden and certain of the priorities have different consequences. All in all, 13 such differential equations result. These are derived in more detail in Appendix B.

The equations for the alternating priority case are derived in exactly the same way, and are also stated in Appendix B. The mixed priority case is somewhat more complex, leading to 17 equations. The additional equations are required because we must distinguish whether j falls in the range 1 to r − 1 or the range r to n − 1.

Standard probabilistic reasoning leads at once to the following useful formulas:

1. Expected number of aircraft in the system at time t

$$= \sum_{i=0}^{m} \sum_{j=0}^{n} (i + j) P_{ij}(t)$$

2. Probability runway idle at time t

$$= R_{0,0}(t)$$

3. Probability that arrival queue be saturated at time t

$$= P_m(t) = \sum_{j=0}^{n} [P_{m,j}(t) + Q_{m,j}(t)]$$

4. Probability that departure queue be saturated at time t

$$= P_n(t) = \sum_{i=0}^{m} (P_{i\,n}(t) + Q_{i\,n}(t))$$

5. Expected number of landing aircraft turned away in time T

$$= \int_0^T P_m(t)\, \lambda'(t)\, dt.$$

Similarly, expected number of take-off aircraft turned away in time T

$$= \int_0^T P_n(t)\, \lambda(t)\, dt$$

This completes the mathematical formulation of the differential difference equations governing the number of aircraft in the system. It is noted that in each case the system of differential equations is a finite, homogeneous, linear system of equations of the first order.

6.6 WAITING TIMES AND DELAY

The definition of *delay* given in Chapter 3, Section 3.5, was couched in general terms: The actual time T' taken by an aircraft in passing through a given part of the air transportation system (e.g., approach until landing at a terminal) minus the time T taken under ideal conditions. On recognizing that both T' and T are in actual fact of the nature of random variables, instead of using T' – T as the measure of delay, the average (expected value) $\bar{T}' - \bar{T}$ was the number introduced. This quantity can in fact be measured by statistical observations at air terminals, and it can also be calculated by analytical tools on the basis of models. We are now in a position to carry out the latter process by use of the methods of the last two sections. But when we do so, we shall find that there are alternative choices for the measure of delay, depending on how, in the mathematical definition of T, we conceive of "ideal conditions." This will lead us to two possible choices, each being meaningful and useful in its own sphere.

The mathematical question is one of *waiting times* in a queue. There are two extremes in the possible assumptions concerning the individual service times: the deterministic one that regards each aircraft of given type as requiring exactly h units of time for service once its turn to land has come; and the most purely random assumption that its service time is a chance variate having a *mean* of h. If the exponential distribution based on the parameter υ is assumed, $h = 1/\upsilon$, as is well known. However, in the former case the *variance* time is 0 whereas in the latter it is $1/\upsilon^2$.

Suppose that our aircraft reaches the air terminal (the point where it is taken under terminal area control) at the time t, at which time there are (n − 1) aircraft in the queue ahead of it, which must all be serviced (allowed to land) before ours is permitted to land. If H_i is the time for servicing the i'th aircraft (equal to h in the deterministic case, a random variable otherwise), the full time to landing of our aircraft is:

$$T = H_1 + H_2 + \ldots + H_n,$$

and the expected value is:

$$\bar{T} = \bar{H}_1 + \bar{H}_2 \ldots + \bar{H}_n = nh.$$

This is n/υ in the exponential case.

As a first step toward establishing the practical significance of this number, we shall calculate its standard deviation σ, where $\sigma^2 = \overline{T^2} - (\bar{T})^2$. We have, since the H_i are mutually independent:

$$\begin{aligned}\overline{T^2} &= \overline{H_1^2} + \ldots + \overline{H_n^2} + 2\bar{H}_1\ \bar{H}_2 + \ldots + 2\bar{H}_{n-1}\bar{H}_n\\ &= (n^2 + n)\, h^2 \text{ in the exponential case}\\ &= n^2\, h^2 \text{ in the deterministic case.}\end{aligned}$$

Consequently, $\sigma^2 = nh^2$ in the former and $\sigma^2 = 0$ in the latter case. Thus in the most random case, $\sigma = h\sqrt{n}$. When n has a large enough value to make the problem of waiting an important one, e.g., between 16 and 25, σ is one-quarter to one-fifth the value of $\bar{T}$: this mean is, therefore, a good indication of what usually happens.

How, then are we to define "delay?" Taking, as has been done at the outset in Chapter 3, Section 3.5, the "ideal" that our aircraft has no rivals for use of the landing strip, we would define the former $\bar{T}$ as that in which n = 1, giving the value h. If $\overline{T'}$ is the less favorable value, when n > 1, the "delay" comes out to be (n − 1)h.

This figure is based on "conditional" probabilities and averages: all, *given* that the number n is known. What may be more to the point is the delay when only the time of arrival, t, is given as known. Then the average must be over all possible values of n, weighted by their probabilities $P_n(t)$, so that

$$\bar{T} = h \sum_{n=1}^{n} n\, P_n(t) = hN_t$$

which (apart from notation) is calculated in Section 6.4, and evaluated, under the assumptions of that section in Equations (6-12) through (6-14). One has but to read N_t off from the graphs of Section 6.3 for the values.

How are we to measure the term "delay" when these formulas show that the time through the system to landing has a predictable expected value — determined by and calculable from the arrival input? In other words, what are we to regard as time through under "ideal" conditions? If these are interpreted to mean no aircraft but ours using the terminal, this implies a serious departure from reality, since it will never occur under the conditions in which delay is of practical significance.*

There are two possible answers. First, we may calculate the effect of fluctuations, e.g., as measured by a standard deviation, and regard a delay as occurring when the mean time for landing is exceeded. Second, we may abandon the word "delay" – with its somewhat subjective overtones to the air traveler – and estimate the quality of air transportation by the *mean service time* . Since this is not only measurable observationally, but mathematically predictable on the basis of any given policy leading to given inputs, it is a useful quantity to air traffic control.

*The situation is comparable to that of establishing a scientific definition of the term "efficiency" of a heat engine: if it is the actual thermal energy converted to useful work divided by what would be converted under "ideal conditions," we get an answer depending on what conditions are regarded as "ideal": the mechanical equivalent of heat (First Law), or the Carnot Cycle (Second Law). How close to reality is our ideal to be?

Even though the second step has been taken and the results given above, it is still useful to take the first and study the standard deviation in various cases, particularly when the value of the number to be landed n is unknown.

If only the time of arrival t is known, the value of $\bar{T}$ is given as $h\bar{N}_t$ as above; whereas $\overline{T^2}$ must now be given by multiplying by $P_n(t)$ the quantity $(n^2 + n)h^2$ obtained earlier, and then summing over n.

$$\overline{T^2} = h^2 \sum_{n=1}^{\infty} (n^2 + n)\, P_n(t) = h^2(\overline{N_t^2} + \bar{N}_t)$$

$$= h^2(\overline{N_t^2} + \bar{N}_t) - h^2(\bar{N}_t)^2$$

$$= h^2\sigma_N^2 + h^2\bar{N}_t .$$

Thus the total variance in the time to landing has been broken up into the variance $h^2\sigma_N^2$ contributed by the uncertainty of the value of the (random) number of aircraft N_t, and the variance in landing time, given that N_t has its mean value $\bar{N}_t$.

Computation, or the graphs of Section 6.3 in the cases assumed there, show the order of σ^2 at various times of day t and input conditions. Thus, for example, in one figure we have $\bar{N}_t = 23$ and $\sigma_N = 2$ so that, if it takes an average of h = 2 minutes to land, the mean time through is 46 minutes, and the standard deviation of actual landing time away from this mean is the square root of $2^2(2^2 + 23)$; i.e., $\sigma = \sqrt{27} = 10.4$ minutes.

So far the actual computations have used the case of Section 6.4 as an example. It is necessary to realize that other models may be more relevant, such as those the mathematical formulations of which are carried out in Section 6.5. If the time for take-off were the same as that of landing, and the discipline were first-come/first-served, the problem could be reduced to that of a single queue. In more realistic cases, we have queues of mixed composition, and the whole matter of calculating the time through the system becomes more complicated, depending, among other things, upon the queue discipline. The analytical tools, however, are of the same type as those in the various illustrations given above: the machinery is most appropriately set in motion in terms of an exact statement of the practical problem of interest.

We conclude with certain generalities. An ideally managed air traffic control system would be one in which the expected landing and take-off waiting times are a minimum. It may be relevant to mention here that in a paper, entitled "The Effect of Queue Discipline on Waiting Time Variance,"[5] it has been shown that when the following two conditions hold:

1. No server sits idle while there are customers waiting to be served.
2. The probability of a busy period of infinite duration is zero.

(and making no assumptions regarding the form of the inputs and the service time distribution) that the mean waiting time is independent of the queue discipline and the variance of the waiting time is a minimum when the customers are served in order of their arrival. Tambouratzis[6] has shown that the variance of waiting times is a maximum when the queue discipline is last-come/first-served.

The above result could be applied to the theory of air traffic control and the subsequent priority disciplines that we have treated in Section 6.4.3 of this chapter if we were to assume that an aircraft from the landing queue has the same runway use distribution as an aircraft from the take-off queve. And, if the minimum time required between two landings or take-offs or between a landing and a take-off (or between a take-off and a landing) were the same. For such a case, if h is the average time to service an aircraft (including the minimum time between landings, etc.) and the service time distribution is exponential, and if $P_n(t)$ denotes the probability of a *total* of n aircraft in the system (landing and take-off queues lumped together), then the expected waiting time would be independent of queue discipline and calculable by the earlier formulas.

ADDENDUM

In view of the importance of time-dependent effects in air traffic control, and in an attempt to make use of all existing methodology taking these effects into account, an extensive examination of the sources was made based on the following reference material:

The 15 Year Index (Operations Research Society of America).

Bibliography of Queuing Theory (Appearing in The Elements of Queuing Theory by T. L. Saaty).

Mathematical Review.

Operations Research – Management Service (Executive Services Institute).

The following five papers appear to be the most relevant for the problems in "Air Traffic Control System Capacity Measurement Methodology":

Kendall, D. G., On the Generalized Birth and Death Process, Annals of Mathematical Statistics, Vol. 19, 1948.

Clarke, A. B., A Waiting Line Process of Markov Type, Annals of Mathematical Statistics, Vol. 27, 1956.

Luchak, G., The Solution of the Single Channel Queueing Equations Characterized by a Time-dependent Poisson Distributed Arrival Rate and a General Class of Holding Times, Operations Research, Vol. 4, 1956.

Luchak, G., Distribution of the Time Required to Reduce to Some Preassigned Level a Single Channel Queue Characterized by a Time-dependent Poisson Distributed Arrival Rate and a General Class of Holding Times, Operations Research, 1957.

Von Sydow, L., Some Aspects on the Variations in Traffic Intensity, Teleteknik, 1958.

REFERENCES

1. Galliher, H.P., and Wheeler, R.C., Nonstationary Queuing Probabilities for Landing Congestion of Aircraft, Operations Research, Vol. 6, No. 2, March-April 1958, p. 264.

2. Ibid.: pp. 165–306

3. Cox, D.R., and Miller, H.D., The Theory of Stochastic Processes (John Wiley, New York, 1965) Chs. 3.11 and 3.12.

4. Ibid.: Ch. 1.c

5. Kingman, J.F.C, The Effect of Queue Discipline on Waiting Time Variance, Proc. Camb. Phil. Soc., Vol. 58, 1962.

6. Tambouratzis, On the Property of the Variance of the Waiting Time of a Queue, J. Appl. Rob., Vol. 3, 1968, pp. 702–703.

7. CONCLUSIONS

The concept of capacity of the air traffic control system is linked unavoidably with that of capacity of the air transportation system as a whole. Many different peak and average flow rates can be identified with capacity. Each use of the word capacity refers to the maximum value of some rate or amount under a particular set of constraints. A change in the constraints may produce a change in the capacity: thus agreement must be reached about the constraints before capacity is defined. When the capacity of the system as a whole is at issue, many of the constraints are thresholds of acceptable service quality, such as acceptable average delay or acceptable risk. In this case the specification of capacity depends not only on the units which are counted and the assumed values of many operating variables, but also on the criteria according to which service is judged satisfactory or unsatisfactory.

Quantitative analysis of capacity is impossible without a concurrent analysis of safety for the following reasons: If no other adaptations were made in the air transportation system, a simple increase in the amount of traffic would probably result in a disproportionate increase in the accident rate. In real life, the air traffic control system responds adaptively to an increase in traffic demand by degrading the service in other less severe ways, such as delays and cancellations, while maintaining safety at a high level and obscuring its connection to capacity. Thus there is no direct relation between amount of traffic and other service degradation, only an indirect connection through the direct connection each has with afety. A number of measures of safety have been defined, including one, the probability of fatality per hour of exposure of the subject, which is particularly relevant in comparing the risk of flying to other socially accepted risks. However, indirect methods of measuring the safety of operating systems must be developed, for accidents are too infrequent to provide a valid basis for many important decisions concerning safety.

To relate a definition or measurement of capacity to an existing or proposed ATC system, we need a canonical description of the system to show how service degradations result from increased traffic. To have predictive value, this description should not be tied to present ATC system implementation. For this reason, we have undertaken to make a description of air traffic control in terms of goals and functions rather than in terms of specific equipment configuration and performance specifications. We have completed a preliminary step and have found no reason to doubt the feasibility of such an effort.

Inasmuch as service degradations are often linked with transient stress and peak load conditions, and because demand and environment may change rapidly, stationary time-invariant queues are an inadequate representation of real air transportation system queues when capacity is under study. We have turned to the mathematical theory of time-varying queues and applied it to problems representative of air traffic control. We have proved the existence of periodic solutions to a large class of queueing problems with periodic driving and service functions, representative of diurnal variations in demand and service rate. Queue statistics such as mean queue length and waiting time, and their respective standard deviations, and expected number of users turned away can be calculated by solving differential equations, without the use of Monte Carlo or other simulation processes. Numerical calculations with a simple single queue example have illustrated time-dependent relations among queue characteristics which could not be discovered from a steady-state queue analysis.

APPENDIX A

THE TYRANNY OF SMALL NUMBERS: PROBLEMS IN THE STATISTICAL ANALYSIS OF AIRCRAFT ACCIDENTS

A.1 DISCUSSION

It is difficult to draw worthwhile conclusions from the statistical analysis of events as infrequent as aircraft accidents. This can be illustrated with some manipulation of the selected data shown in Tables A-1 and A-2.

Table A-1 shows some selected fatal accident figures[1,2] involving U.S. certified route air carrier scheduled passenger service. This is our safest class of service, and includes only a very small number of fatal accidents: 78 fatal accidents in the period 1956 to 1967 inclusive. For each of these years we have tabulated the number of fatal accidents, the revenue miles flown, the revenue hours flown, and the number of departures. Fatal accidents in which the only fatalities were to occupants of another aircraft which was not a certified air carrier in scheduled passenger service are excluded. We have also tabulated the totals for the 6-year periods 1956 to 1961 inclusive and 1962 to 1967 inclusive, as well as the 12-year grand totals.

Table A-2 shows similar figures for U.S. general aviation flying, one of our more dangerous classes of service. An estimate of the number of departures was not readily available, so these data are omitted. Because of the large number of fatal accidents it will easily result that statistically va conclusions fall out freely.

What model of the probability of occurrence of a fatal accident should we use? Under the assumption that these are extremely rare events, virtually independent of one another, we would expect the number to be a sample from a Poisson distribution.

We now pose the question: Are the flights of these vehicles becoming safer with the passage of time? Before that question has a meaning, we must define what we mean by safety. We have already discussed the choice of units for the estimation of safety, and we shall try out three different normalizations: number of fatal accidents per mile flown, number of fatal accidents per hour of flight, and number of fatal accidents per departure.

TABLE A-1

SELECTED FATAL ACCIDENT FIGURES
U.S. CERTIFIED ROUTE AIR CARRIER SCHEDULED PASSENGER SERVICE

Year	No. Fatal Accidents	Revenue Miles Flown (100 millions)	Revenue Hours Flown (100 thous.)	No. of Departures (100 thous.)
1956	4	8.4	39.1	34.5
1957	5	9.5	43.2	37.2
1958	6	9.5	42.7	36.1
1959	10	10.1	44.4	38.9
1960	10	9.8	40.2	38.3
1961	5	9.6	36.0	37.3
1962	5	10.0	34.6	36.5
1963	5	10.8	35.5	37.7
1964	9	11.7	37.1	39.3
1965	7	13.4	40.1	41.8
1966	4	14.7	42.9	43.5
1967	8	18.1	48.5	49.1
1956-1961	40	56.9	245.6	222.3
1962-1967	38	78.7	238.7	247.9
Total	78	135.6	484.3	470.2

TABLE A-2

SELECTED FATAL ACCIDENT FIGURES
U.S. GENERAL AVIATION FLYING

Year	No. Fatal Accidents	Miles Flown (100 millions)	Hours Flown (100 thous.)
1956*	356	13.1	102.0
1957	438	14.3	109.4
1958*	384	16.6	125.7
1959*	450	17.2	129.0
1960*	429	17.7	131.2
1961*	426	18.6	136.0
1962*	430	19.6	145.0
1963*	482	20.5	151.1
1964	504	21.8	157.4
1965	538	25.6	167.3
1966	573	33.4	210.2
1967	576	34.4	221.5
1956-1961	2,483	97.5	733.3
1962-1967	3,103	155.3	1,052.5
Total	5,586	252.8	1,785.8

*Est.

To make our statistical manipulations easier to grasp, we shall perform the analysis graphically. Our medium is binomial probability paper, as described in reference 3. The ordinate scales lay out length proportional to the square root of the numerical value of the respective variables. For graphical representation, this system of scales has two great advantages: (1) the magnitude of the standard deviation in distance units is the same on any part of the graph; and (2) the skewness of a Poisson distribution is largely balanced by the non-linearity of the scale.

In Figure A-1, the number of fatal accidents is plotted against revenue miles flown for U.S. certified route air carriers in scheduled passenger traffic for each of the years 1956 through 1967. A solid line through the origin represents the mean rate of approximately 5.85 fatal accidents per billion miles flown. Parallel to this line are drawn dashed lines at intervals of one and two standard deviations above and below the mean rate. The vertical line segments are plots of the data points from Table A-1. For technical reasons (discussed in reference 3, page 206) each is plotted as a vertical line segment extending from the datum number to the next higher integer instead of as a point. The horizontal scale is revenue miles flown in units of 10 million. It is obvious that these data cluster around the line representing the mean rate. The distribution of displacements of these 12 points from the mean rate appears entirely consistent with the assumption that the expected mean rate is the same for all 12 years. No points are as far as two standard deviations from the overall mean rate, and only three are more than one standard deviation away (the location of each, for this purpose, is the center of the line segment).

In Figure A-2 a similar plot of the number of fatal accidents is presented as a function of revenue hours flown, also for U.S.-certified route air carriers in scheduled passenger traffic. Once again, the spread of points is not larger than one would expect from a random selection of Poisson distributions having rates defined by the overall mean rate. Similarly, in Figure A-3, the number of fatal accidents is plotted against the number of departures for U.S.-certified route air carriers in scheduled passenger traffic. For a third time, the distribution is not manifestly non-random.

The picture is different when we consider general aviation. Figure A-4 shows the number of fatal accidents plotted as a function of miles flown for general aviation over the same 12-year period. Because the number of accidents is large, a different scale,in which the vertical line segments degenerate into points, has been selected. On this scale, the standard deviation is much

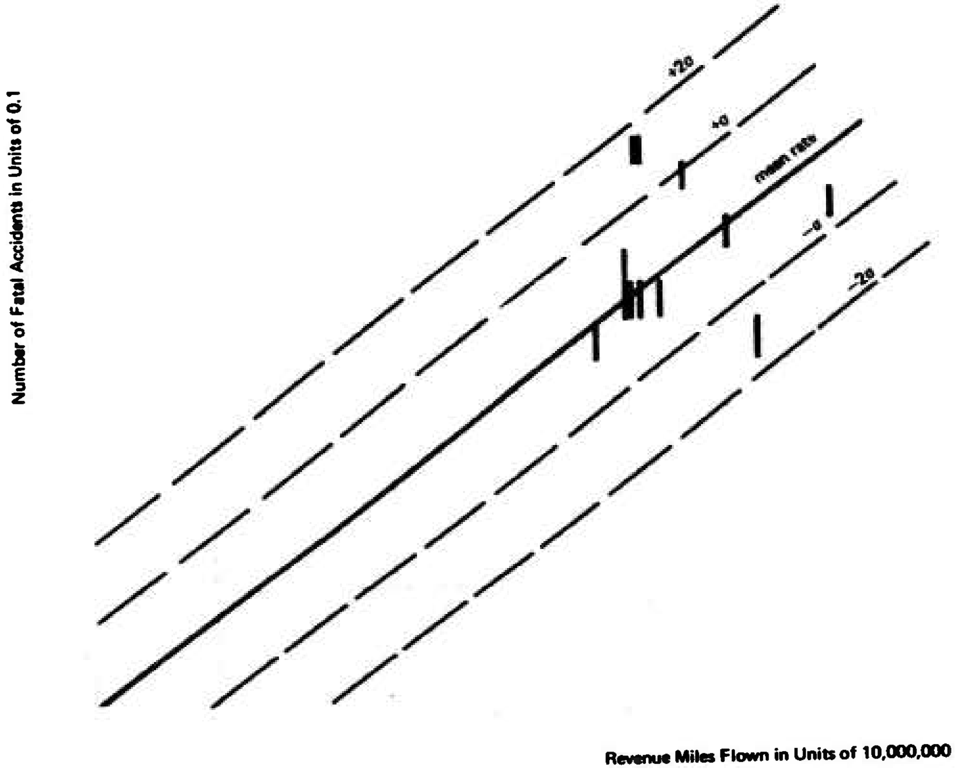


FIGURE A-1 NUMBER OF FATAL ACCIDENTS VERSUS REVENUE MILES FLOWN—U.S. CERTIFIED ROUTE AIR CARRIERS, SCHEDULED PASSENGER TRAFFIC, 1956-1967

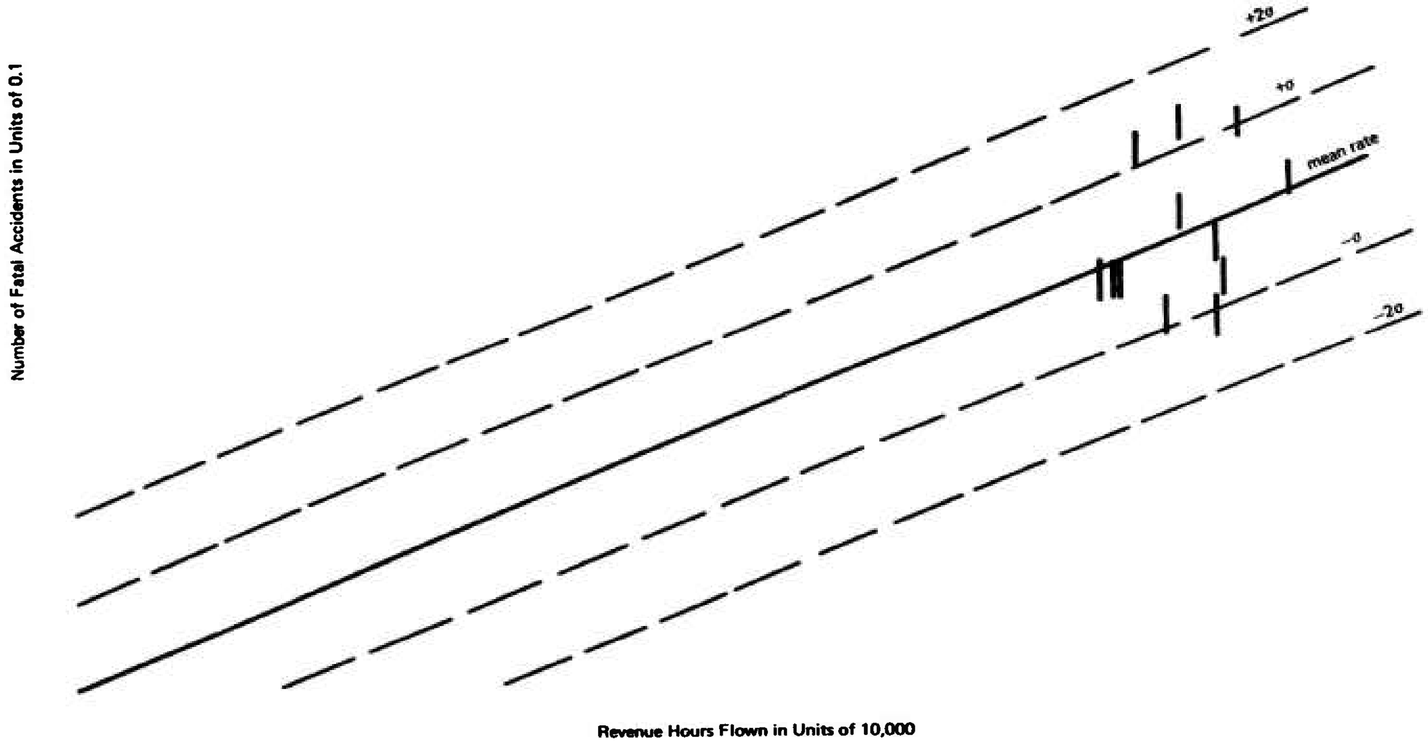


FIGURE A-2 NUMBER OF FATAL ACCIDENTS VERSUS REVENUE HOURS FLOWN—U.S. CERTIFIED ROUTE AIR CARRIERS, SCHEDULED PASSENGER TRAFFIC, 1956-1967

NOT REPRODUCIBLE

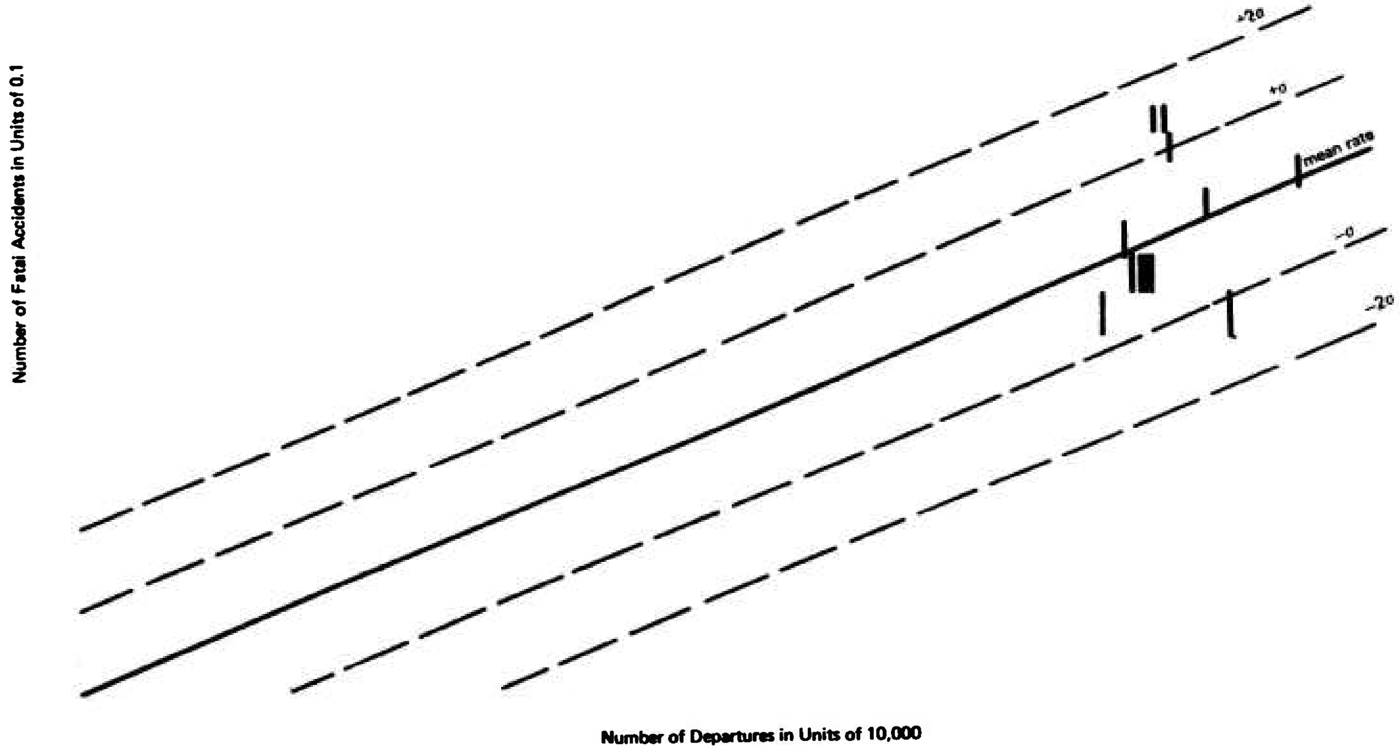


FIGURE A-3 NUMBER OF FATAL ACCIDENTS VERSUS NUMBER OF DEPARTURES—U.S. CERTIFIED ROUTE AIR CARRIERS, SCHEDULED PASSENGER TRAFFIC, 1956-1967

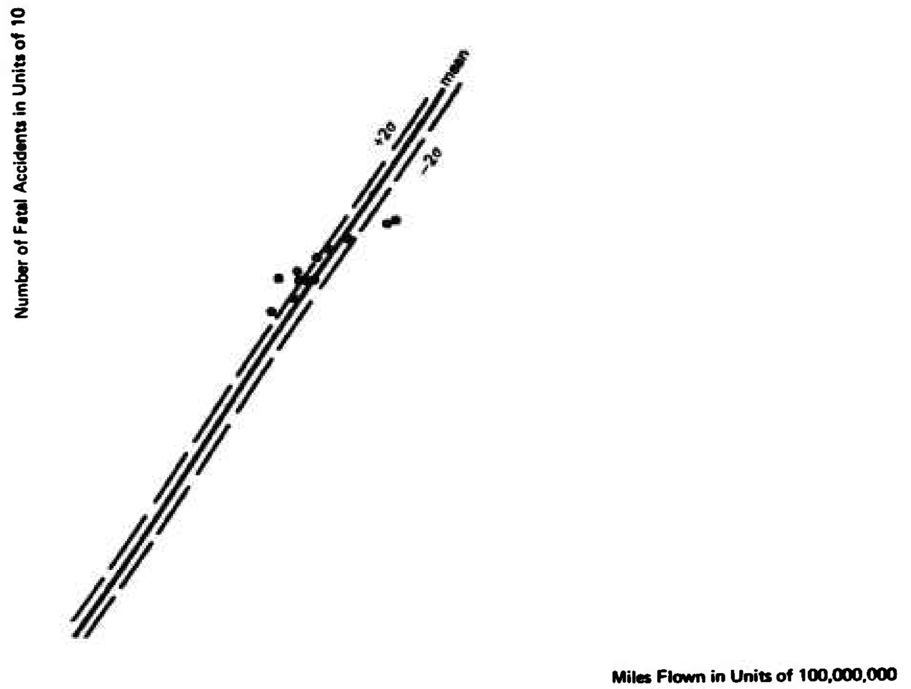


FIGURE A-4 NUMBER OF FATAL ACCIDENTS VERSUS MILES FLOWN–U.S. GENERAL AVIATION, 1956-1967 smaller, and the only guidelines shown are two standard deviations above and below the mean value. With 5 points out of 12 more than two standard deviations from the mean, this group of data is manifestly not a random selection from Poisson distributions with means defined by the solid curve. Inasmuch as the number of general aviation miles flown per year is an increasing function of the year over this period of time, we can read the points chronologically from left to right. This shows, furthermore, that there is a definite trend toward a lower rate of fatal accidents per mile over this 12-year period.

Figure A-5 shows a similar plot of the number of fatal accidents versus hours flown by U.S. general aviation in the same 12-year period. Here, only two points fall more than two standard deviations from the mean, and only four or perhaps five mcre than one standard deviation from the mean. This particular graphical analysis is somewhat inconclusive, but there is good reason to believe that a more refined statistical analysis of the same data would reveal a trend toward a decreasing number of fatal accidents per hour flown. However, if we took only the last 6 years (1962 through 1967), such a statement could not be made.

If comparing the data for scheduled passenger service on a year-by-year basis is inconclusive, what happens if we group the data into larger bundles? On Figure A-6 the 12-year period is divided into two periods, 1956 through 1961 and 1962 through 1967, and the same data as in Figures A-1, -2, and -3 are replotted. The number of fatal accidents per mile is less in the later period than in the earlier, and the difference is statistically significant at a confidence level of ab^ut 5 percent. On the other hand, although the number of fatal accidents per departure also decreases, the two values are not statistically distinguishable. This supports a prediction made in 1962 by Fromm.[4]

Finally, the rate of fatal accidents per hour of travel shows no decrease from one 6-year period to the next. This confirms the observation of Starr[5] that the commercial aviation accident rate per hour of exposure appears to be stabilizing.

Figure A-7 shows a similar plot of the fatal accident rates of U.S. general aviation in terms of hours or miles. In both instances, there is a statistically significant decrease in the accident rate. However, one can argue whether the numerical decrease from 33.8 to 29.5 fatal accidents per million flying hours is meaningful in other senses, even if it can be supported statistically.

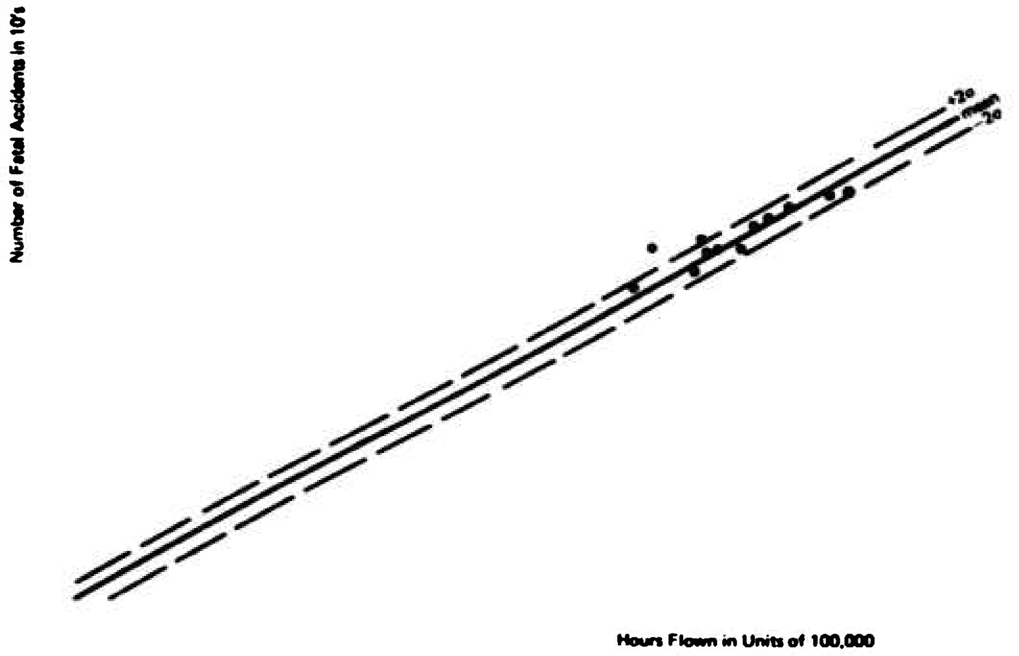


FIGURE A-5 NUMBER OF FATAL ACCIDENTS VERSUS HOURS FLOWN—U.S. GENERAL AVIATION, 1956-1967

NOT REPRODUCIBLE

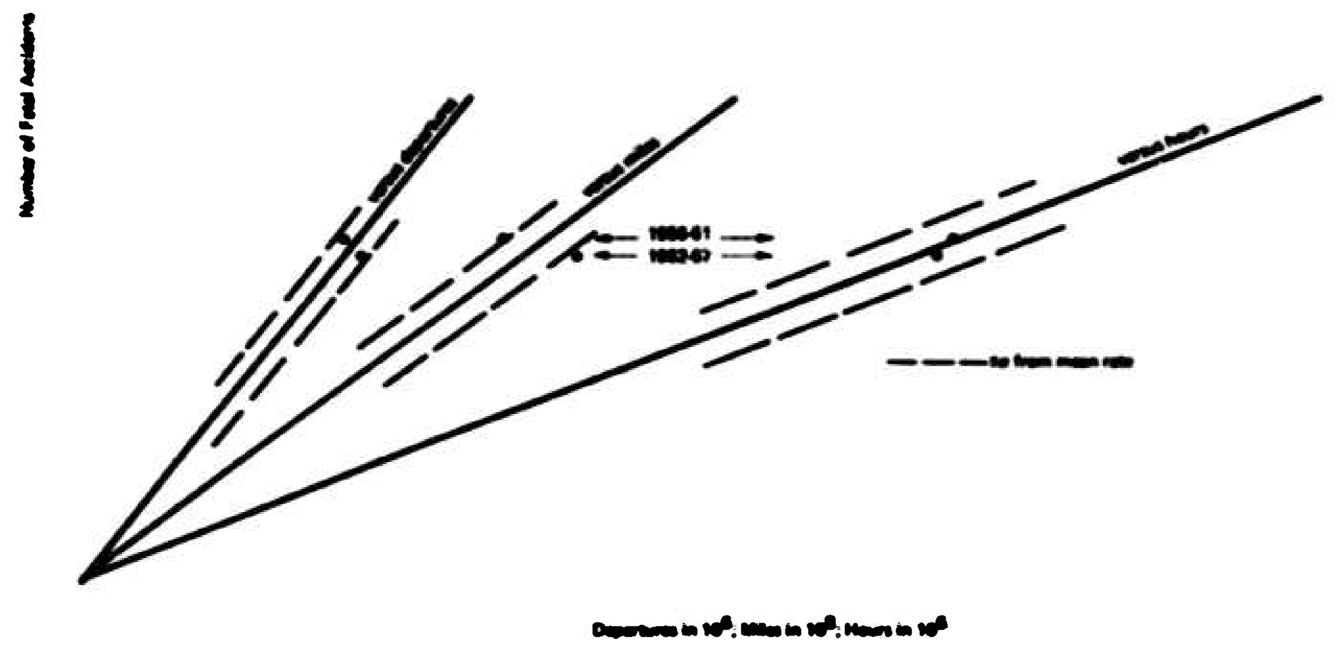


FIGURE A-6 COMPARISON OF FATAL ACCIDENT RATES DURING 1958-61 AND 1962-67, U.S. CERTIFIED ROUTE AIR CARRIERS, SCHEDULED PASSENGER SERVICE

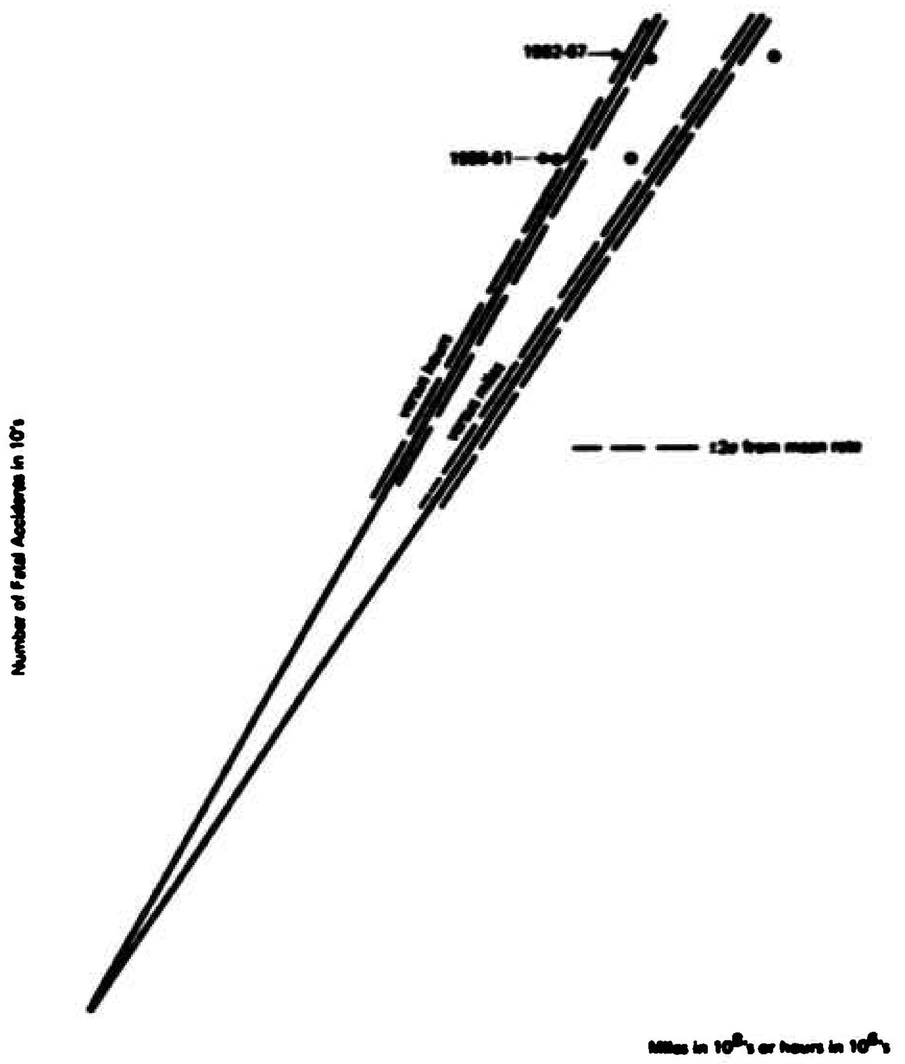


FIGURE A-7 COMPARISON OF FATAL ACCIDENT RATES DURING 1958-61 AND 1962-67, U.S. GENERAL AVIATION

The same kind of analysis will allow us to do some elementary design of experiments. Suppose we wish to gather data for the purpose of establishing as statistically valid a supposed reduction in the rate of occurrence of accidents (or any other discrete events). Suppose, as before, that these events are sparse and statistically independent, occurring at some average rate per unit of exposure. The exposure unit may be miles, hours, thousands of departures, or any like quantity related to the physical mechanisms and operations of the system under study. Suppose that the existence of a mean rate has been established by copious measurements in the past and its value is known. How many observations must we make to show that a new and reduced value of the mean exists?

The answer to this question depends on the value of the new mean. If the new mean is close to the old mean, any observations are required; whereas if the new mean is far removed from the old mean, fewer observations suffice.

We answer this question graphically by drawing, on binomial probability paper again, a straight line through the origin representing the established mean. The abscissa represents exposure in arbitrary units; the ordinate represents number of events observed. Then, using the scale printed on the upper left-hand portion of the paper, we draw lines parallel to the established mean line displaced downward at distances of 1.29, 2.33, and 3.09 standard deviations. These are the arguments for which the standard probability integral has the values 0.900, 0.990, and 0.999 respectively. Then, we draw a line representing the new mean. On Figure A-8, two illustrative cases have been drawn, one with a new mean half as great as the established mean, and one with a new mean three-quarters as great as the established mean. The intersections of the new mean lines with the three lines previously drawn show, approximately, the number of events which must be observed to establish, with confidence of 10, 1, or 0.1 percent (one-sided) that a point lying on the new mean line differs in a statistically significant way from the old established mean.

Thus, to show with 99 percent confidence that some change has cut the accident rate in half, we must carry on enough experiments to observe eight accidents under the new regime. To show that the accident rate has been reduced 25 percent, with 99 percent confidence, we must gather data on at least 60 events at the new rate.

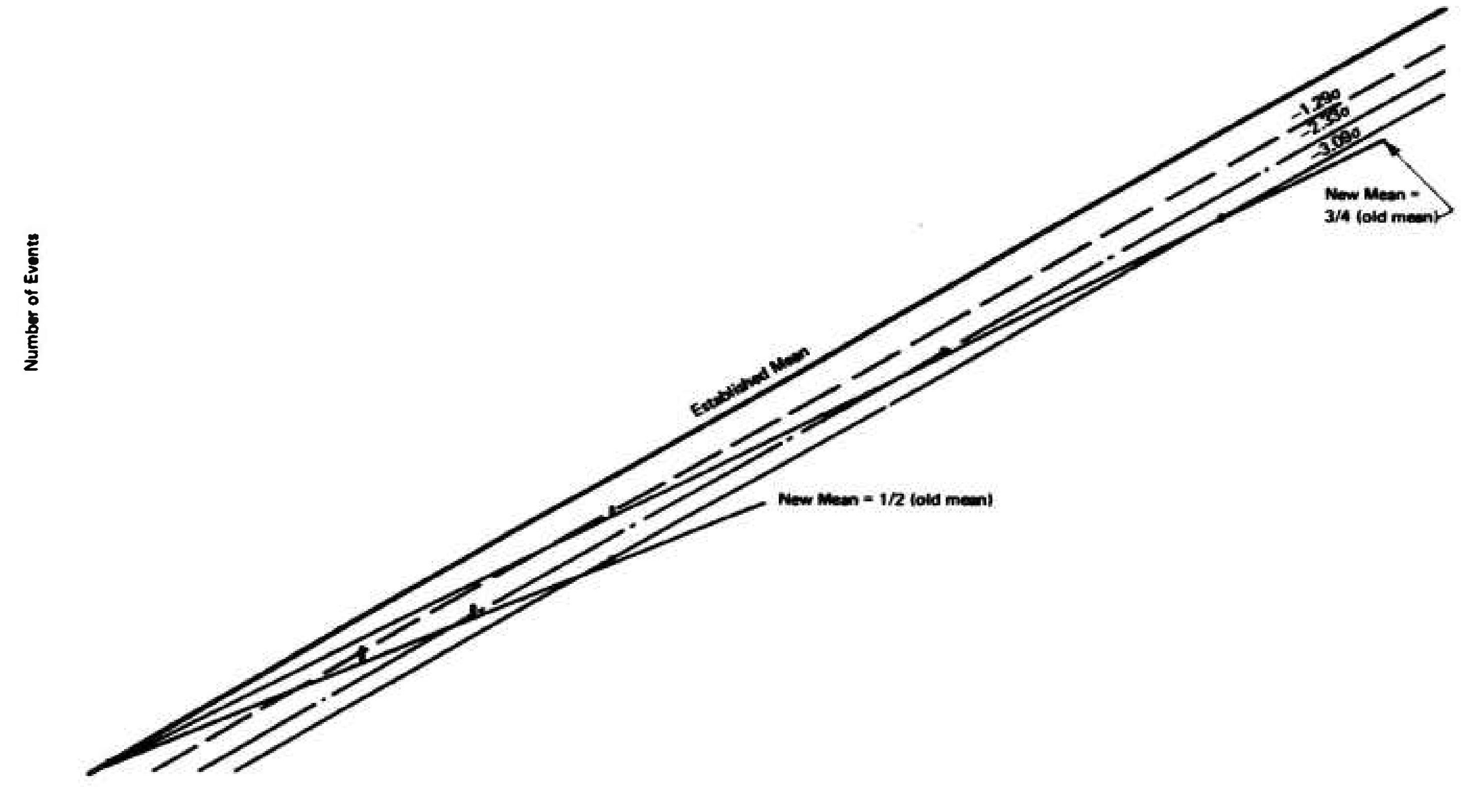


FIGURE A-8 DESIGN OF EXPERIMENT TO PROVE REDUCTION IN THE RATE OF A POISSON PROCESS

A.2 CONCLUSIONS

From the illustrated examples above, we can draw the following conclusions.

- At the rate at which accidents occur in scheduled passenger service of certified domestic airlines, the number of events occurring in a few years is too small for us to draw non-trivial, statistically valid conclusions about the degree of improvement in safety. At the rate at which fatal accidents occur in general aviation, however, non-trivial conclusions are possible.

- A secondary set of conclusions, incidental to the purpose of this report and arising simply from the illustrative examples, can also be drawn, as follows. The rate of fatal accidents per plane mile in domestic scheduled passenger service in the years 1962 through 1967 is significantly smaller than the same rate in the period 1956 through 1961. However, if the safety is normalized in units of fatal accidents per departure or fatal accidents per revenue hour of operation, no decrease can be demonstrated. In the case of general aviation, the decrease in fatal accident rate is statistically significant whether measured in miles of exposure or hours of exposure, although the amount of the decrease is numerically small in the latter units.

REFERENCES

1. *FAA Statistical Handbook of Aviation,* 1967 Edition, Department of Transportation, Federal Aviation Administration, November 1967.

2. *FAA Statistical Handbook of Aviation,* 1968 Edition, Department of Transportation, Federal Aviation Administration, 1968.

3. Mosteller, F., and Tukey, J. W., *The Uses and Usefulness of Binomial Probability Paper,* J. Am. Stat. Assn., Vol. 44, June 1949, pp. 174–212.

4. Fromm, G., *Economic Criteria for Federal Aviation Agency Expenditures,* Final Report, Prepared for Federal Aviation Agency (Contract No. FAA/BRD-355), United Research Incorporated, June 1962.

5. Starr, C., Social Benefit versus Technological Risk, *Science,* Vol. 165, No. 3899, September 19, 1969, pp. 1232–1238.

APPENDIX B

DETAILS OF THE MATHEMATICAL FORMULATION OF THE INTERACTION OF LANDING AND TAKE-OFF QUEUES UNDER TIME-DEPENDENT CONDITIONS

This appendix is intended to be read in conjunction with Section 6.5 of Chapter 6, and the notation and assumptions of that chapter are followed here and will not be restated.

At that point, we derived the following transient equation for the state probability.

$$\begin{aligned} P_{i,j}(t + \Delta t) &= P_{i,j}(t)\ [1 - \lambda(t)\Delta t - \lambda'(t)\Delta t - \mu'_{i,j}\ \Delta t] \\ &\quad + P_{i-1,j}(t)\ \lambda'(t)\Delta t + P_{i,j-1}(t)\ \lambda(t)\ \Delta t \\ &\quad + P_{i+1,j}(t)\ \mu'_{i+1,j} + Q_{i,j+1}(t)\ \mu_{i,j+1}\ \Delta t \quad \text{for} \quad \begin{array}{l} 0 < i < m \\ 0 < j < m \end{array} \end{aligned}$$

If either i or j has one of its extreme values, certain of the transitions are forbidden. We can imagine the diagram of Figure 6-12 reduced and placed at the appropriate point on Figure 6-11. If it lands on a boundary, one transition is forbidden, and if it lands on a corner, two of these transitions are forbidden. The corresponding terms in the equations above must be appropriately modified. The results are as follows:

$$\begin{aligned} P_{i,0}(t + \Delta t) &= P_{i,0}(t)\ [1 - \lambda(t)\Delta t - \lambda'(t)\Delta t - \mu'_{i,0}\Delta t] \\ &\quad + P_{i-1,0}(t)\ \lambda'(t)\Delta t + P_{i+1,0}(t)\ \mu_{i+1,0}\ \Delta t \quad \text{for} \quad 0 < i < m \\ &\quad + Q_{i,1}(t)\ \mu_{i,1}\ \Delta t \end{aligned}$$

$$\begin{aligned} P_{m,0}(t + \Delta t) &= P_{m,0}(t)\ [1 - \lambda(t)\Delta t - \mu_{m,0}\ \Delta t] + P_{m-1,0}(t)\ \lambda'(t)\ \Delta t \\ &\quad + Q_{m,1}(t)\ \mu_{m,1}\ \Delta t \end{aligned}$$

$$\begin{aligned}
P_{m,j}(t + \Delta t) &= P_{m,j}(t)\ [1 - \lambda(t)\Delta t - \mu'_{m,j}\Delta t] + P_{m-1,j}(t)\lambda'(t)\ \Delta t \\
&\quad + P_{m,j-1}(t)\ \lambda(t)\Delta t + Q_{m,j+1}(t)\ \mu_{m,j+1}\ \Delta t; \qquad 0 < j < m
\end{aligned}$$

$$\begin{aligned}
P_{i,n}(t + \Delta t) &= P_{i,n}(t)\ [1 - \lambda'(t) - \mu'_{i,n}\ \Delta t) + P_{i-1,n}(t)\ \lambda'(t)\ \Delta t \\
&\quad + P_{i,n-1}(t)\ \lambda(t)\ \Delta t + P_{i+1,n}(t)\ \mu'_{i+1,n}\ \Delta t \qquad 0 < i < m
\end{aligned}$$

$$\begin{aligned}
P_{m,n}(t + \Delta t) &= P_{m,n}(t)\ [1 - \mu'_{m,n}\ \Delta t] + P_{m-1,n}(t)\ \lambda'(t)\ \Delta t \\
&\quad + P_{m,n-1}(t)\ \lambda(t)\ \Delta t
\end{aligned}$$

$$\begin{aligned}
R_{0,0}\ t + \Delta t &= R_{0,0}(t)\ [1 - \lambda(t)\ \Delta t - \lambda'(t)\ \Delta t] + Q_{0,1}(t)\ \mu_{0,1}\ \Delta t \\
&\quad + P_{1,0}(t)\ \mu'_{1,0}\ \Delta t
\end{aligned}$$

By the same process, we enumerate the ways to arrive at state $Q_{i,j}(t + \Delta t)$, with the following results:

$$\begin{aligned}
Q_{i,j}(t + \Delta t) &= Q_{i,j}(t)\ [1 - [\lambda(t) + \lambda'(t) + \mu_{i,j}\ \Delta t]] + Q_{i,j-1}(t)\ \lambda(t)\ \Delta t \\
&\quad + Q_{i-1,j}(t)\ \lambda'(t)\ \Delta t \qquad \begin{array}{l} 0 < i < m \\ 0 < j < n \end{array}
\end{aligned}$$

$$\begin{aligned}
Q_{0,j}(t + \Delta t) &= Q_{0,j}(t)\ [1 - \lambda(t)\ \Delta t - \lambda'(t)\ \Delta t - \mu_{0,j}\Delta t] + Q_{0,j-1}(t)\ \lambda(t)\ \Delta t \\
&\quad + P_{1,j}(t)\ \mu'_{1,j}\ \Delta t + Q_{0,j+1}(t)\ \mu_{0,j+1}\ \Delta t - \text{ for } 0 < j < n
\end{aligned}$$

$$Q_{0,n}\,(t + \Delta t) = Q_{0,n}(t)\,[1 - \lambda'(t)\,\Delta t - \mu_{0,n}\,\Delta t] + Q_{0,n-1}(t)\,\lambda(t)\,\Delta t$$

$$+ P_{1,n}(t)\,\mu'_{1,n}\,\Delta t$$

$$Q_{m,j}(t + \Delta t) = Q_{m,j}(t)\,[1 - [\lambda(t) + \mu_{m,j}\Delta t]] + Q_{m-1,j}(t)\,\lambda'(t)$$

$$+ Q_{m,j-1}(t)\lambda(t)\,\Delta t \qquad 0 < j < n$$

$$Q_{i,n}(t + \Delta t) = Q_{i,n}(t)\,[1 - [\lambda'(t) + \mu_{i,n}\Delta t]]$$

$$+ Q_{i-1,n}(t)\,\lambda'(t)\,\Delta t + Q_{i,n-1}(t)\,\lambda(t)\,\Delta t \qquad 0 < i < m$$

$$Q_{m,n}(t + \Delta t) = Q_{m,n}(t)\,[1 - \mu_{m,n}\,\Delta t] + Q_{m-1,n}(t)\,\lambda'(t)\Delta t + Q_{m,n-1}(t)\,\lambda(t)\,\Delta t$$

For completeness, we may add the nondefined states:

$$P_{o,j}(t + \Delta t) \equiv 0; \;\; 1 \leqslant j \leqslant n$$

$$Q_{i,0}(t) = 0 \qquad 0 < i \leqq m$$

On multiplying, transposing $P_{i,j}(t)$, $(0 < i < m,\ 0 < j < n)$ to the left, and dividing by Δt, the system of equations becomes:

$$\frac{P_{i,j}(t + \Delta t) - P_{i,j}(t)}{\Delta t} = -\,[\lambda(t) + \lambda'(t) + \mu'_{i,j}]\;P_{i,j}(t)$$

$$+ P_{i-1,j}(t)\,\lambda'(t) + P_{i,j-1}(t)\,\lambda(t) + P_{i+1,j}(t)\,\mu'_{i+1,j}$$

etc., etc.

If we take limits as Δt→0, then, by definition, the left side is the derivative $dP_{ij}(t)/dt$, etc., and the equations become

$$\begin{aligned}\frac{d}{dt} P_{i,j}(t) = & - [\lambda(t) + \lambda'(t) + \mu'_{i,j}]\; P_{i,j}(t) + \lambda'(t)\, P_{i-1,j}(t) \\ & + \lambda(t)\, P_{i,j-1}(t) + \mu'_{i+1,j}\; P_{i+1,j}(t) \qquad 0 < i < m \\ & + \mu_{i,j+1}\; Q_{i,j+1}(t) \qquad 0 < j < n\end{aligned} \tag{B-1}$$

$$\begin{aligned}\frac{d}{dt} P_{i,0}(t) = & - [\lambda(t) + \lambda'(t) + \mu'_{i,0}]\; P_{i,0}(t) + \lambda'(t)\, P_{i-1,0}(t) \\ & + \mu'_{i+1,0}\; P_{i+1,0}(t) + \mu_{i,1} Q_{i,1}(t) \qquad 0 < i < m\end{aligned} \tag{B-2}$$

$$\begin{aligned}\frac{d}{dt} Q_{0,j}(t) = & - [\lambda(t) + \lambda'(t) + \mu_{0,j}]\; P_{0,j}(t) + \lambda(t)\, Q_{0,j-1}(t) \\ & + \mu'_{1,j}\; P_{1,j}(t) + \mu_{0,j+1}\, Q_{0,j+1}(t) \qquad 0 < j < n\end{aligned} \tag{B-3}$$

$$P_{0,j}(t) = Q_{i,0}(t) = 0 \qquad 0 < j \leqslant m \qquad 0 < i \leqslant n$$

$$\frac{d}{dt} P_{m,0}(t) = - [\lambda(t) + \mu'_{m,0}]\; P_{m,0}(t) + \lambda'(t)\, P_{m-1,0}(t) + \mu_{m,1} Q_{m,1}(t) \tag{B-4}$$

$$\frac{d}{dt} Q_{0,n}(t) = - [\lambda(t) + \mu_{0,n}]\; P_{0,n}(t) + \lambda(t)\, Q_{0,n-1}(t) + \mu'_{1,n} P_{1,n}(t) \tag{B-5}$$

$$\begin{aligned}\frac{d}{dt} P_{m,j}(t) = & - [\lambda(t) + \mu'_{m,j}]\; P_{m,j}(t) + \lambda'(t)\, P_{m-1,j}(t) \\ & + \lambda(t)\, P_{m,j-1}(t) + \mu_{m,j+1}\; Q_{m,j+1}(t) \qquad 0 < j < m\end{aligned} \tag{B-6}$$

$$\frac{d}{dt} P_{i,n}(t) = -[\lambda'(t) + \mu'_{i,n}] P_{i,n}(t) + \lambda'(t) P_{i-1,n}(t) \qquad (B\text{-}7)$$

$$+ \lambda(t) P_{i,n-1}(t) + \mu_{i+1,n} P_{i+1,n}(t) \qquad 0 < i < m$$

$$\frac{d}{dt} P_{m,n}(t) = -\mu'_{m,n} P_{m,n}(t) + \lambda'(t) P_{m-1,n}(t) + \lambda(t) P_{m,n\text{-}1}(t) \qquad (B\text{-}8)$$

$$\frac{d}{dt} R_{0,0}(t) = -[\lambda(t) + \lambda'(t)] R_{0,0}(t) + \mu_{0,1} Q_{0,1}(t) + \mu'_{1,0} P_{1,0}(t)^{*} \qquad (B\text{-}9)$$

$$\frac{d}{dt} Q_{i,j}(t) = -[\lambda(t) + \lambda'(t) + \mu_{i,j}] Q_{i,j}(t) + \lambda(t) Q_{i,j-1}(t) \qquad (B\text{-}10)$$

$$0 < i < m$$

$$+ \lambda'(t) Q_{i-1,j}(t) \qquad 0 < j < n$$

$$\frac{d}{dt} Q_{m,j}(t) = -[\lambda(t) + \leftarrow \mu_{m,j}] Q_{m,j} + \lambda(t) Q_{m-1,j}(t) \qquad (B\text{-}11)$$

$$+ \lambda'(t) Q_{m,j-1}(t)$$

$$\frac{d}{dt} Q_{i,n}(t) = -[\leftarrow \lambda'(t) + \mu_{i,n}] Q_{i,n}(t) + \lambda'(t) Q_{i-1,n}(t) \qquad (B\text{-}12)$$

$$+ \lambda(t) Q_{i,n-1}(t) \qquad 0 < i < m$$

$$\frac{d}{dt} Q_{m,n}(t) = -\mu_{m,n} Q_{m,n}(t) + \lambda'(t) Q_{m-1,n}(t) + \lambda(t) Q_{m,n-1}(t) \qquad (B\text{-}13)$$

*$R_{0,0}(t)$ = probability that there are no aircraft in either queue.

B 1 ALTERNATING PRIORITIES

The equation for this and the next (mixed priority) case is based on reasoning similar to the preceding. The reader should have no difficulty in supplying the details.

$$\begin{aligned}\frac{d}{dt} Q_{i,j}(t) &= -\,[\lambda(t) + \lambda'(t) + \mu_{i,j}]\, Q_{i,j} + Q_{i-1,j}(t)\lambda'(t) + Q_{i,j-1}(t)\lambda(t)\\ &\quad + P_{i+1,j}(t)\, \mu'_{i+1,j} \qquad 0 < i < m \qquad 0 < j < n\end{aligned}$$

$$\begin{aligned}\frac{d}{dt} P_{i,j}(t) &= -\,[\lambda(t) + \lambda'(t) + \mu'_{i,j}]\; P_{i,j}(t) + Q_{i,j+1}(t)\, \mu_{i,j+1}\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad 0 < i < m\\ &\quad + P_{i-1,j}(t)\, \lambda'(t) + P_{i,j-1}(t)\, \lambda(t) \qquad\qquad 0 < j < n\end{aligned}$$

$$\begin{aligned}\frac{d}{dt} Q_{0,j}(t) &= -\,[\lambda(t) + \lambda'(t) + \mu_{0,j}]\; Q_{0,j}(t) + P_{1,j}(t)\;\; \mu'_{1,j}\\ &\quad + Q_{0,j+1}(t)\, \mu_{0,j+1}(t) + \lambda(t)\, Q_{0,j-1}(t) \qquad 0 < j < n\end{aligned}$$

$$\begin{aligned}\frac{d}{dt} P_{i,0}(t) &= -\,[\lambda(t) + \lambda'(t) + \mu'_{i,0}]\; P_{i,0}(t) + P_{i+1,0}(t)\, \mu'_{i+1,0}\\ &\quad + Q_{i,1}(t)\, \mu_{i,1} + P_{i-1,0}(t)\, \lambda'(t) \qquad\qquad 0 < i < m\end{aligned}$$

$$\begin{aligned}\frac{d}{dt} P_{m,0}(t) &= -\,[\lambda(t) + \mu'_{m,0}]\; P_{m,0}(t) + P_{m-1,0}(t)\, \lambda'(t)\\ &\quad + Q_{m,1}(t)\, \mu_{m,1}\end{aligned}$$

$$\frac{d}{dt} Q_{0,n}(t) = -[\lambda'(t) + \mu_{0,n}] Q_{0,n}(t) + Q_{0,n-1}(t) \lambda(t) + P_{1,n}(t) \mu'_{1,n}$$

$$\begin{aligned} \frac{d}{dt} P_{m,j}(t) &= -[\lambda(t) + \mu'_{m,j}] P_{m,j}(t) + P_{m-1,j} \lambda'(t) \\ &\quad + P_{m,j-1}(t) \lambda(t) + Q_{m,j+1}(t) \mu_{m,j+1} \qquad 0 < j < n \end{aligned}$$

$$\begin{aligned} \frac{d}{dt} Q_{m,j}(t) &= -[\lambda(t) + \mu_{m,j}] Q_{m,j}(t) + Q_{m-1,j}(t) \lambda'(t) \\ &\quad + Q_{m,j-1}(t) \lambda(t) + P_{m+1,j}(t) \mu'_{m+1,j} \qquad 0 < j < n \end{aligned}$$

$$\begin{aligned} \frac{d}{dt} Q_{i,n}(t) &= -[\lambda'(t) + \mu_{i,n}] Q_{i,n}(t) + Q_{i-1,n}(t) \lambda'(t) \\ &\quad + Q_{i,n-1}(t) \lambda(t) + P_{i+1,n}(t) \mu'_{i+1,n} \qquad 0 < i < m \end{aligned}$$

$$\begin{aligned} \frac{d}{dt} P_{i,n}(t) &= -[\lambda'(t) + \mu'_{i,n}] P_{i,n}(t) + P_{i-1,n}(t) \lambda'(t) \qquad 0 < i < m \\ &\quad + P_{i,n-1}(t) \lambda(t) + Q_{i,n+1}(t) \mu_{i,n+1} \end{aligned}$$

$$\frac{d}{dt} P_{m,n}(t) = \mu'_{m,n} P_{m,n}(t) + P_{m-1,n} \lambda'(t) + P_{m,n-1} \lambda(t)$$

$$\frac{d}{dt} Q_{m,n}(t) = -\mu_{m,n} Q_{m,n}(t) + Q_{m-1,n}(t) \lambda'(t) + Q_{m,n-1}(t) \lambda(t)$$

$$\frac{d}{dt} R_{0,0}(t) = -[\lambda(t) + \lambda'(t)] R_{0,0}(t) + P_{1,0}(t) \mu'_{1,0} + \mu_{0,1} Q_{0,1}(t)$$

It is convenient to add a definition:

$R_{i,j}(t)$ $\equiv$ unconditional probability of there being i aircraft in the landing queue and j in the take-off queue.

$= P_{i,j}(t) + Q_{i,j}(t) \qquad 0 < i < m \qquad 0 < j < n$

$R_{i,0}(t) = P_{i,0}(t) \qquad P_{0,j}(t) = Q_{i,0}(t) = 0$

$R_{0,j}(t) = Q_{0,j}(t)$

This is consistent with the use of $R_{0,0}(t)$ as the probability that both queues have zero length and k is undefined.

B.2 MIXED PRIORITIES

For mixed priorities, the equations are:

$$\frac{d}{dt} P_{i,j}(t) = -[\lambda(t) + \mu'_{i,j} + \lambda'(t)] P_{i,j}(t) + \lambda'(t) P_{i-1,j}(t) + \lambda(t) P_{i,j-1}(t) + \mu'_{i+1,j} P_{i+1,j}(t) + \mu_{i,j+1} Q_{i,j+1}(t) \qquad \begin{array}{l} 0 < i < m \\ 0 < j < r \end{array}$$

$$\frac{d}{dt} P_{i,j}(t) = -[\lambda(t) + \lambda'(t) + \mu'_{i,j}] P_{i,j}(t) + \lambda'(t) P_{i-1,j}(t) + \lambda(t) P_{i,j-1}(t) \qquad \begin{array}{l} 0 < i < m \\ r < j < n \end{array}$$

$$\frac{d}{dt} Q_{i,j}(t) = -[\lambda(t) + \lambda'(t) + \mu_{i,j}] Q_{i,j}(t) + \lambda(t) Q_{i,j-1}(t) + \lambda'(t) Q_{i-1,j}(t) \qquad \begin{array}{l} 0 < i < m \\ 0 < j < r \end{array}$$

$$\frac{d}{dt} Q_{i,j}(t) = - [\lambda(t) + \lambda'(t) + \mu_{i,j}] Q_{i,j}(t) + \lambda(t) Q_{i,j-1}(t)$$

$$+ \lambda'(t) Q_{i-1,j}(t) + \mu_{i,j+1} Q_{i,j+1}(t) + P_{i+1,j}(t) \mu'_{i+1,j} \qquad \begin{array}{c} 0 < i < m \\ r \leq j < n \end{array}$$

$$\frac{d}{dt} P_{i,0}(t) = -[\lambda(t) + \lambda'(t) + \mu'_{i,0}] P_{i,0}(t) + \lambda'(t) P_{i-1,0}(t)$$

$$+ \mu'_{i+1,0} P_{i+1,0}(t) + \mu_{i,1} Q_{i,1}(t) \qquad 0 < i < m$$

$$\frac{d}{dt} Q_{0,j}(t) = - [\lambda(t) + \lambda'(t) + \mu_{0,j}] P_{0,j}(t) + \lambda(t) Q_{0,j-1}(t)$$

$$+ \mu'_{i,j} P_{i,j}(t) + \mu_{0,j+1} Q_{0,j+1}(t) \qquad 0 < j < n$$

$Q_{0,j}$is exceptional – there is no aircraft in the arrival queue, and so a plane in the departure queue is served, no matter whether $j<r$ or $j \geq r$.

$$\frac{d}{dt} P_{m,0}(t) = -[\lambda(t) + \mu'_{m,0}] P_{m,0}(t) + \lambda'(t) P_{m-1,0}(t) + \mu_{m,1} Q_{m,1}(t)$$

$$\frac{d}{dt} Q_{0,n}(t) = -[\lambda'(t) + \mu_{0,n}] P_{0,n}(t) + \lambda(t) Q_{0,n-1}(t) + \mu'_{1,n} P_{1,n}(t)$$

$$\frac{d}{dt} P_{m,j}(t) = - [\lambda(t) + \mu'_{m,j}] P_{m,j}(t) + \lambda'(t) P_{m-1,j}(t) + \lambda(t) P_{m,j-1}(t)$$

$$+ \mu_{m,j+1} Q_{m,j+1}(t); \qquad 0 < j < n$$

$$\frac{d}{dt} P_{m,j}(t) = -[\lambda(t) + \mu'_{m,j}] P_{m,j}(t) + \lambda'(t) P_{m-1,j}(t) + \lambda(t) P_{m,j-1}(t)$$

$$r \leqslant j \leqslant n$$

$$\frac{d}{dt} P_{i,n}(t) = -[\lambda'(t) + \mu'_{1,n}] P_{i,n}(t) + \lambda'(t) P_{i-1,n}(t) + \lambda(t) P_{i,n-1}(t)$$

$$\frac{d}{dt} P_{m,n}(t) = -\mu'_{m,n} P_{m,n}(t) + \lambda'(t) P_{m-1,n}(t) + \lambda(t) P_{m,n-1}(t)$$

$$\frac{d}{dt} Q_{m,j}(t) = -[\lambda(t) + \lambda'(t) + \mu_{m,j}] Q_{m,j}(t) + \lambda'(t) Q_{m-1,j}(t) + \lambda(t) Q_{m,j-1}(t)$$

$$0 < j < r$$

$$\frac{d}{dt} Q_{m,j}(t) = -[\lambda(t) + \lambda'(t) + \mu_{m,j}] Q_{m,j}(t) + \lambda'(t) Q_{m-1,j}(t) + \lambda(t) Q_{m,j-1}(t)$$

$$+ \mu_{m,j+1} Q_{m,j+1}(t) \qquad r \leqslant j < n$$

$$\frac{d}{dt} Q_{m,n}(t) = -\mu_{m,n} Q_{m,n}(t) + \lambda'(t) Q_{m-1,n}(t) + \lambda(t) Q_{m,n-1}(t)$$

$$\frac{d}{dt} Q_{i,n}(t) = -[\lambda(t) + \lambda'(t) + \mu_{i,n}] Q_{i,n}(t) + \lambda'(t) Q_{i-1,n}(t) + \lambda(t) Q_{i,n-1}(t)$$

$$0 < i < m$$

$$\frac{d}{dt} R_{0,0}(t) = -[\lambda(t) + \lambda'(t)] R_{0,0}(t) + \mu_{0,1} Q_{0,1}(t) + \mu'_{1,0} P_{1,0}(t)$$

$R_{0,0}(t)$ = probability there are no aircraft in either queue.

Clearly $P_{0,j}(t) = Q_{i,0}(t) = 0$ per $0 < j \leqslant m$, $0 < i \leqslant n$.